Topic Representation Learning on Sequential Data for Text Understanding

by

Lixing Zhu

Thesis

Submitted to the University of Warwick

in partial fulfilment of the requirements

for admission to the degree of

Doctor of Philosophy in Computer Science

Department of Computer Science

The University of Warwick

March 2023

Contents

List of Tables

List of Figures

Acknowledgements

Pursuing a PhD degree has never been an easy process, and I am fortunate to have been supported by so many wonderful people. I would foremost like to thank my supervisor Prof. Yulan He who supported me with countless hours of discussions, valuable advice, ideas and encouragement. Without you I would not have gone this far. Thank you for giving me the freedom to pursue diverse research directions and being patient in guiding me through so many ups and downs. I would also like to thank my second supervisor Dr. Gabriele Pergola without whose inspiration I would not have taken the journey leading to this work. Dr. Lin Gui gave me valuable supervision early in my PhD, from whom I learned much. I would also like to thank the CS Department staff, especially Andrew and Richard, for their help and suggestions throughout my PhD study. I also learned a great deal from brilliant collaborators: Zheng, Runcong and others. There are many others I have not collaborated with but have nevertheless contributed to my work-and-life time: Junru, Akanni, John, Xingwei, Hanqi, Wenjia, Jiazheng, Jun, Xinyu and Zhaoyue. There are many people who have made my PhD journey memorable: Ouyang, Shikai, Yiming, Yueqi, Zongpai, Quirin, Shan, Vassan and all the others. Many thanks also go to the advisors for their insightful comments and feedback. The journey is not at the end. The road ever goes on. Finally, I am eternally grateful to my girlfriend, Maple Wei, not only for backing me and encouraging me at all times, but also for her sparkling ideas and enthusiasm for life. And to my parents: Thank you for enduring my absence over the past 3 years and supporting me all the time. You have always been there whenever I was set back, giving me a solid foundation based on your unconditional love. Without you I would not have been where I am today.

Declarations

The work in this thesis was conducted by the author during the period September 2019 - January 2023 at the University of Warwick, in collaboration with Prof. Yulan He. Chapter 3 generalises some arguments from our work in [332] and Chapter 4 uses results from [331]. Chapters 5 and 6 are based on the findings from [333, 334]. Where we make use of work not our own or rework established arguments, we write (for instance): "we follow Pergola et al. [210]". To the best of my knowledge, the material contained in this thesis is original and my own work except where otherwise stated. This thesis has not been submitted for a degree at any other university.

Abstract

Neural sequence models have become prevalent owing to the sequential nature of natural language and high expressiveness of neural networks. Despite achieving a huge success, however, such frameworks are challenged for their ineptness in capturing the global context or compressing holistic features such as style and topic, nor could they disentangle latent representation into factors of interest and informative variables.

In this thesis, we build models to learn topic representations capturing the thematic feature as well as develop semi-supervised learning techniques to exploit the inductive bias from few annotated data. We introduce (1) topic representation learning via fine-tuning of denoising auto-encoders that fits topic modelling into a seq2seq structure; (2) aspect-stance disentanglement using constrained priors that improves classification of vaccination stance and text spans; (3) disentangled cross attention to inject inductive bias of different dimensions with different objectives; and (4) swapping auto-encoder that promotes the instance-level discrimination for aspect-stance disentanglement in order to perform clustering along different latent factor dimensions. Besides, a vaccination attitude dataset containing tweets about Covid vaccines is constructed for the validation of the proposed approaches.

We provide empirical studies of the proposed models, showing that topic representation acquired by fine-tuning language models is opportune for capturing the latent semantics. More importantly, with few annotations, such representations can be disentangled under the constraint of additional prior or by disentangled cross attention, significantly improving the performance of stance classification and aspect span detection. By incorporating the siamese network that forms the swapping auto-encoder, we are able to cluster tweets along the axis of aspects that has been successfully disentangled.

Chapter 1

Introduction

Throughout history, texts have been essential in communication between humans and machines, as well as among humans themselves. Texts are used for multiple purposes, such as expressing emotions to human beings, providing instructions to machines, and transmitting knowledge across generations. Language is a mirror of mind [48]. The thoughts transmitted by texts are extraordinary complex, yet it is surprisingly simple that texts are predominantly token sequences or strings. The ability to comprehend text in a linear fashion is not only innate to humans, but it is also theoretically advantageous, if not the most effective method for machines. This is evidenced by the utilization of paper tape I/O Turing Machines [274]. Therefore, the development of automated text comprehension systems holds significant potential, particularly in scenarios where only unstructured data are available [11, 18, 47, 262, 277]. The focus of this thesis is on the modelling of sequential data using representation learning approaches in order to facilitate the understanding of semantics, with applications in text classification and clustering.

1.1 Motivation

Text sequences are ubiquitous [275]. The sequential nature of languages gave rise to myriads of sequence models [18], playing a fundamental role in Natural Language Processing (NLP). The latter, which are better known as Sequence-to-Sequence (Seq2Seq) models, has achieved great success across a range of NLP tasks, e.g., Speech Recognition [82, 119], Handwriting Recognition [81], Machine Translation [116], Text Generation [293] and Language Modelling [18, 34], to name a few. Due to the sequential structure inherent in natural languages, neural sequence models such as LSTMs and Transformers [296] are well-suited for capturing the complex relationships between tokens, sentences, paragraphs, and discourses.

Despite the tremendous success in various tasks, Seq2Seq models are less efficient in capturing the global features or high-level properties such as style and topic [27], when compared with latent variable models or Bayesian nonparametric approaches [31]. While the volume of online documents continues to grow, the inability of models to comprehend vast amounts of unlabelled data has become an aggravating problem [56, 226], compounded by the rapid increase of parameters [14, 128, 167, 197]. It has been demonstrated that end-to-end learning easily fits randomly generated training data despite the increase of parameters [312]. If such a model is applied to text understanding, we will find it difficult to navigate across domains and adapt to different styles, themes, or contexts. For example, in a Vaccination Corpus [189], "*rash*" commonly refers to a symptom or disease. Conversely, in TV sitcoms (e.g., The Big Bang Theory)[1] [42], the characters are often reminded by "*Don't make any rash decisions*". In this situation, the model will be easily confused by the switch of the domains. While this circumstance can be mitigated by fine-tuning word embedding models (e.g., pre-trained language models) [213] or semi-supervised learning [99], the holistic properties of the local context, scilicet the tone, topic and syntactic style at the sentence level [31], cannot be appropriately captured. For instance, tweets are geared towards different topics, and the TV transcripts are rendered with different emotions in different scenarios. On the contrary, Bayesian models, e.g., the topic models, have a more principled way to leverage the statistics of co-occurrences, uncovering the distributional property of the latent topics in an unsupervised manner [150], with less parametric redundancy [3]. Although there are variational recurrent neural networks [51] designed for the generalization of the local context, such structures have not been fully exploited in the era of pre-training and fine-tuning. In this thesis, we aim to leverage the expressiveness of neural sequence models and generalisation of Auto-Encoding Variational Inference networks [123] to learn both the global and local contextual information at different levels from text, bringing them together into a semi-supervised framework that can be used in a variety of NLP applications.

Aside from capturing the co-occurrence patterns, latent variable models show advantages in exhibiting a certain level of disentanglement and interpretability —– capturing human-understandable characteristics from data [118]. For example, a person can distinguish negative vaccine-related tweets from a collection of negative grocery reviews, as long as they recognize vaccines and food, even though they have

[1] `https://bigbangtrans.wordpress.com/series-7-episode-24-the-status-quo-combustion/`

never seen such training instances before. The abilities to generate new samples and steer controllable factors are especially valued in Text Generation [243, 260] and Text Style Transfer [115], where human initiatives are desired to create out-of-distribution (OOD) data. These abilities also matter in sentiment analysis applications where people often reach out to convey their experiences and seek emotional support [330]. If we take the dialogue system of GPT-3 as an example, a human interlocutor will receive the following responses[2]: '*Q: What is your favorite animal? A: My favourite animal is dog. Q: Why? A: Because dogs are loyal and friendly.*'. Although the response of the agent seems plausible and coincides well with the commonsense knowledge, it is less natural compared with '*I once raised a dog in my childhood, he is my best friend.*'. It is often the case that the empathy and initiatives of human beings which brings vibrancy to life are nonexistent in machines. While Seq2Seq models excel in fitting the dependency, latent variable models are more effective interpolating between the inferred hidden semantics, as reflected in recent generative frameworks, such as Variational Auto-Encoding Bayes (VAE) [31, 123, 140] or Generative Adversarial Nets (GAN) [78, 170, 215], in generating controlled text [66] or stylish images [204]. In this thesis, we delve into the intersection of Seq2Seq and latent variable models, bridging the gap between prediction and interpolation. While there are multitudes of work targeting text generation from a well-designed latent space [53] (e.g., hyperbolic space), we introduce disentangled learning to factor out independent components [109] such as stance and aspect. By extrapolating the pre-trained model and disentangling the latent space, we hope that the structured semantics will facilitate attitude detection in online posts.

Another desirable property we expect the sequential models to have is the ability to cluster. After all, the notion of category does not spawn out of thin air, nor are they provided by human annotations [134]. Recent years have witnessed a surge of pre-trained language models in unsupervised self-learning for natural language processing, yet their full potential of detecting emerging categories in the context of online documents remains unexploited. As a motivating example, consider the two tweets "*If you're worried about the blood clot, do not read the leaflet in a box of Paracetamol*!" and "*There are some very interesting ties between this vaccines creators and the eugenics movement which is concerning considering it's mainly been promoted as a vaccine for poor folks in the third world.*" The first tweet ironically addressed vaccine side effects and the second one expressed instead specific political concerns. This is different from traditional aspect-based sentiment analysis on product reviews where a small number of exhaustive aspects are pre-defined. Traditional

[2] https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html

classifiers built upon sequence-to-sequence architectures [11, 47] and attention mechanisms [210, 305] fall short in detecting unseen points [14]. While recent progress in NLP resorts to the '*Human in the Loop*' [337] protocol or '*Text span prediction*' [185] scheme to accommodate this low-resource setting, they are of less practical use for open-domain tasks such as clustering and information retrieval, where a unidimensional vector for semantic similarity measure is preferred [225]. In contrast, denoising auto-encoder [98, 287] learns a bottleneck representation [167, 187] that interpolates between data manifolds in its fine-tuning. Such representations can afford similarity measures in clustering algorithms which boost the performance. More importantly, the idea of hidden representation learning leads to the application of disentanglement [19, 125, 172], in line with the intuition that humans categorize text from independent perspectives. For example, the tweets *"mRNA vaccines are poison"* and *"The Pfizer vaccine is safe"* are both targeting safety issues, whilst they manifest the opposite stances. Most approaches will be obfuscated by the entangled semantics, evidenced by clustering over stance rather than aspect. However, given the presumption that these factors are independent components composing the outward features, even few training samples can be generalized into prominent biases [188]. To this end, we propose to learn disentangled representations that are clustering-friendly in different dimensions and beneficial to downstream tasks. We hope that the disentangled representations are elementary factors that generalise well to unseen identities, even though there are few annotations available.

In the rest of this section, we first introduce the research objectives of this thesis, followed by contributions towards these objectives. Finally we overview each chapter and list the outline.

1.2 Research Objectives

The central theme of this thesis is situated in the intersection of sequential models and topic representation learning, with their applications to sentiment analysis and text clustering. Sequence-to-sequence models provide powerful expressiveness, fitting complex relationship between text sequences, while overparameterised in its nature [14]. Topic modelling enforces a level of formality or compression to the sequential models [27], enjoys better generalisation and disentanglement of latent features [167], and can be easily tweaked to learn clustering-friendly representations [187]. With topic modelling, it is possible to derive holistic semantics from any sentence [140, 187], enabling the disentanglement of latent representation into primitive factors [172]. The latter would facilitate text clustering in the desired semantic space, and naturally allows the recombination of the factors through cross-attention, which increases generalisation and improves the performance in the low-resource tasks. Specifically, we focus on the Sentiment Analysis [253] task since sentiment is the epitome of perceptible factors expressed in languages. A good representation learning method of sentiment and its associated factors would be easily transferred to other NLP tasks such as Text Generation [2, 292], where a certain level of diversity and interpretability is desired. Apart from sentiment analysis, we aim to show the feasibility of disentangled learning in sequential modelling by applying the disentangled representations to text span detection and text clustering. These are two tasks curated in the low-data regime where human annotations are expensive, time-consuming, and may face ethical issues. Recent advances in neural sequence models barely addressed the data scarcity problem and out-of-distribution prediction. In contrast, we assume that topic modelling captures holistic statistics while sequential models encode the sequential dependencies. On this basis, we propose to perform disentangled learning by modifying the architecture of the pre-trained models.

The sequence-to-sequence relationships we are targeting are output-output relationships and output-input relationships, depending on the data format, i.e., whether it is free-form texts or multiple-choice categories according to the LLaMA taxonomy [270]. Among different levels of context, we primarily focus on sentence-level dependency, e.g., the dependency between utterances in a user timeline or conversation thread, using token-level representations of language models as atomic representations. On top of sequence-to-sequence relationships, we aim to build topic representation learning approaches that gauge sentence semantics. Ideally, topic representations are low-dimensional embeddings that distil desirable properties from collocations of words. Given the holistic implicature and condensation of LM's token-level semantics, it is reasonable to expect that such bottleneck representations enable semantic search based on distance metrics, thereby allowing clustering in the semantic space.

Aside from methodology innovations, we also work on dataset construction to fuel the model with inductive biases that complement the knowledge transferred by pre-trained language models. The introduction of the dataset can not only serve as a test bed for the proposed model but also facilitate cross-domain adaptation. In brief, the research objectives (ROs) we define in this thesis are:

RO 1 Modelling the intricate dependencies between different levels of text. Modeling the dependencies of text sequences has always been a key step in text understanding, and different NLP tasks require the adaptation of various building blocks tailored for diverse objectives. In the area of social media analysis, we need to analyse communication flow, quantify the social influence and predict opinion dynamics. In dialogue emotion detection, we make predictions based on the historical conversational context. Despite the variety of task objectives, we believe that sequence models are able to capture the fundamental dependencies among textual components. Consequently, the sequence-to-sequence framework will be customised to incorporate domain-specific features, which is one of the key objectives of this thesis.

RO 2 **Learning topic representations that gauge the global context and capture thematic properties of sequential data.** Topic representations are assumed to encode the co-occurrence patterns within the local context, as well as efficiently generalize the domain-specific statistics of the entire corpus. On the other hand, sequence-to-sequence models tend to make accurate predictions given abundant data labelled. We thereby plan to follow the semi-supervised learning framework to combine the advantages of the two by first performing unsupervised topic modelling and then training task-specific classifiers. With topic modelling, the model will capture the local context and global context of sentences more efficiently, adding generalisability to the classification heads.

RO 3 **Disentangling latent factors from unstructured text using sequential models.** Disentangled learning allows the model to factor out variables of variation associated with observational changes. Hence, the model could recombine or sample from the latent space for the generation of novel data, which in return increases its effectiveness in fitting new data points or interpolating between data manifolds. To this end, the aim is to learn disentangled representations from sufficient new data and a limited amount of manual annotations, which are clustering-friendly and will presumably improve classification results.

RO 4 **Evaluation of topic modelling and disentangled learning on downstream tasks.** We posit that the integration of sequence-to-sequence architecture, topic representation learning and disentangled learning will capture latent semantics characterising both holistic features and compositional patterns. On this provisos, we need to evaluate how the acquired representations redeem the coherent semantics (w.r.t. human evaluations), and the extent to which this learning process improved performance on downstream tasks. We choose sentiment analysis and text clustering since both sentiment and clusters can be topic-dependent features, in addition to word representation learning that requires cross-domain polysemy.

RO 5 **Dataset curation.** In order to be able to evaluate the generalisation capability of our proposed framework, we intend to create datasets covering multiple topics. The datasets will be made up of massive unannotated documents and a handful of annotated instances to facilitate the evaluation of unsupervised topic acquisition and disentangled learning in the low-resource setting. For sentiment analysis, annotations are provided as aspect labels, aspect spans, or stance polarities depending on the sub-task. For text clustering, we need to provide an aspect label or argumentative pattern for each cluster as the groundtruth.

1.3 Contributions

The work of this thesis is situated in the field of NLP addressing the research objectives by jointly learning disentangled representations and training sequence-to-sequence classifiers, under the semi-supervised framework. The major contributions can be summarized as follows:

C. 1 We propose a novel generative model, namely JTW, to jointly learn topics and topic-specific word embeddings. The model leverages both local co-occurrence patterns and global topic distributions to derive contextualised meanings of words. The generative process can also be applied to documents represented by pre-trained language models to endow words with topic-dependent meanings. The obtained word representation better captures word semantics in terms of word similarity evaluation and word sense disambiguation, and the extracted topics are semantically more coherent.

C. 2 We develop a neural temporal opinion model for the prediction of opinion dynamics on Twitter taking into account both the temporal relation and user context by means of sequence-to-sequence prediction and topic modelling. We experimented on two Twitter datasets to show the benefits yielded by the above method.

C. 3 We target the refinement of auto-encoders. We propose topic-driven fine-tuning by inserting a topic layer into a language model whose representations are acquired during the unsupervised training of a variational recurrent attention network. The topic layer captures the conversational topics and tones which are subsequently applied to the sequence-to-sequence prediction of dialogue emotions. Moreover, we incorporate external knowledge from ATOMIC by either SBERT-based extraction or COMET-based generation. We perform empirical analysis to show its effectiveness.

C. 4 We consider the disentanglement of independent latent variables. We design a semi-supervised framework, called VADet, for disentangled aspect/stance representation learning and aspect span detection on tweet corpora. This model, comprising both unsupervised topic representation learning and supervised aspect-stance disentanglement, employs a denoising variational auto-encoder to learn topic representations and uses a constraint on prior to induce the disentanglement. We build a dataset which relates to vaccine attitude detection to afford fine-tuning on in-domain corpus and supervised training with inductive biases and provide extensive evaluations on the proposed dataset.

C. 5 We explore the disentanglement of aspect and stance semantics in the task of text clustering, where we exploits both denoising auto-encoder for topic acquisition and inductive biases for clustering-friendly representation learning. We adopt a swapping-auto encoder and devise a disentangled cross attention to improve the disentanglement between aspect and stance. The proposed method is evaluated on two Covid-19 vaccination corpora with various distance metrics for text clustering, the result of which confirms that disentangled representations substantially improve the performance of clustering algorithms.

1.4 Publications

The work in this thesis is anchored in the following articles and publications, listed in ascending order according to the year of publication.

- **Lixing Zhu**, Yulan He, Deyu Zhou. *"A neural generative model for joint learning topics and topic-specific word embeddings"*. Transactions of the Association for Computational Linguistics (TACL), 2020.
- **Lixing Zhu**, Yulan He, Deyu Zhou. *"Neural Temporal Opinion Modelling for Opinion Prediction on Twitter"*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020.
- **Lixing Zhu**, Gabriele Pergola, Lin Gui, Deyu Zhou, Yulan He. *"Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection"*. In Proceedings of the 59th Annual Meeting of the Association for Computational

Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL), 2021.

- **Lixing Zhu**, Zheng Fang, Gabriele Pergola, Rob Procter, Yulan He. *"Disentangled Learning of Stance and Aspect Topics for Vaccine Attitude Detection in Social Media"*. In Proceedings of The North American Chapter of the Association for Computational Linguistics (NAACL) conference, 2022.

- **Lixing Zhu**, Runcong Zhao, Gabriele Pergola, Yulan He. *"Disentangling Aspect and Stance via a Siamese Autoencoder for Aspect Clustering of Vaccination Opinions"*, Findings of the Association for Computational Linguistics: ACL 2023.

Co-authored publications are written in collaboration with other researchers during the development of this thesis, but they do not form part of the thesis:

- Runcong Zhao, Miguel Arana Catania, **Lixing Zhu**, Elena Kochkina, Lin Gui, Arkaitz Zubiaga, Rob Procter, Maria Liataka, Yulan He. *"PANACEA: An automated misinformation detection system on COVID-19"*. In Proceedings of the 17th conference of the European Chapter of the Association for Computational Linguistics (EACL), 2023

1.5 Thesis Outline

Chapter 1 explains the research area and motivations, along with an overview of the proposed methodologies.

Chapter 2 reviews the literature relevant to topic representation learning and sequential modelling on text understanding. The methods span across prototypical sequence-to-sequence models (e.g., BiLSTM, Transformer), neural topic models, aspect-based sentiment classification models, disentangled learning methods, text clustering methods and semi-supervised approaches. Their relevance to the proposed methodologies concludes each section of this chapter.

Chapter 3 elaborates the joint learning of topics and topic-dependent word embeddings, where observed tokens are taken as generated from topic representations. Topics are inferred by a variational auto-encoder which allows a quick glimpse of the entire corpus. Word representations can be encoded to topic distributions to indicate multiple meanings of words. This chapter is based on the published work of Zhu et al. 2020.

Chapter 4 introduces a sequence-to-sequence model that integrates both user neighbourhood context and tweet stream time interval for stance prediction. We employ a bunch of attention mechanisms to aggregate user timelines and neighbourhood context, which are shown able to improve the performance.

Chapter 5 presents a Seq2Seq model to detect emotions in dialogues, which is composed of a topic layer inserted into a language model whose representations are fine-tuned on downstream datasets. Then each utterance is linked to a phrasal description of the commonsense knowledge such as the reaction of the subject. Finally, a transformer is applied to map a conversation (i.e., an utterance sequence) to an emotion-label sequence. The results are shown to improve on several dialogue emotion detection benchmarks.

Chapter 6 focuses on the task of vaccine attitude detection where the vaccination aspects are unknown and their semantics is entangled with stance. To alleviate the data scarcity problem, a vaccine attitude dataset was constructed from Covid-19 tweets and text span annotations, where the text span indicates the discussed aspect. Chapter 6 also involves the development of a vaccination attitude detection model whose hidden representations are trained in a semi-supervised paradigm. Firstly, part of the model, scilicet the denoising auto-encoder, is trained on large amounts of unannotated tweets to learn latent topics via masked Language Model (LM) learning. Then the model is fine-tuned on a small amount of Twitter data annotated with stance labels and aspect text spans for simultaneous stance classification and aspect span start/end position detection. The model promotes disentanglement in the latent space by putting a constraint on the variational prior and introducing inductive bias from annotations.

Chapter 7 describes a siamese neural network which combines methods of disentangled learning (i.e., disentangled cross attention and swapping auto-encoder) and clustering-friendly representation learning (i.e., denoising auto-encoder) to afford open-domain attitude detection. The latent semantics which is entangled is obtained from unsupervised training of a denoising auto-encoder, whose network weights are retained and subsequently fine-tuned on pair-wise annotated instances. The model enables the clustering algorithms to cluster in a particular semantic space such as aspects based on the distance metrics such as Euclidean distance. Its effectiveness has been demonstrated by empirical results, both quantitatively and qualitatively.

Chapter 8 concludes the thesis and casts insights into future text understanding research by providing new challenges or proposing new directions.

Chapter 2

Literature Review

Chapter Abstract

In this chapter, we first review the neural sequence models and neural topic models that form the basis of the proposed approaches, then we carry on with NLP applications on sentiment analysis and text clustering that our methods are designed for. We begin with sequence-to-sequence learning related to methods presented in Chapter 4, 5 and 6, along with advancements such as RNN Encoder-Decoder and Transformers. Then we move on to latent variable models for topic modelling. After that we proceed with disentangled learning in conjunction with Chapter 6 and 7, which is followed by concepts of semi-supervised learning. Finally, we conclude this chapter with applications to sentiment analysis and text clustering.

2.1 Neural Sequence Models

Neural Sequence Models map an input sequence $\{x_1, x_2, \ldots, x_N\}$ to an output sequence $\{y_1, y_2, \ldots, y_T\}$ by optimizing the joint probability of the output sequence: $\prod_{t=1}^{T} p(y_t|\boldsymbol{x}, y_{<t})$, $\prod_{t=1}^{T} p(y_t|\boldsymbol{x}, y_{-t})$ or $\prod_{t=1}^{T} p(y_t|\boldsymbol{x})$ [18, 47, 277], where the first format is referred to as the autoregressive model in the case $\boldsymbol{y} = \boldsymbol{x}$ [34] and the second is commonly known as the masked language model [60, 67]. Nowadays this framework has enjoyed great success since a wide spectrum of NLP tasks can be formulated as predicting the next label based on the consumption of input and previously generated labels [54], as evidenced by RNN language models [180], named entity recognition (NER) [212], machine translation [262] and speech recognition [82]. The initial

Seq2Seq paradigm comprises a single Recurrent Neural Network (RNN) to carry out sequence labelling tasks [83] such as speech recognition, where labels are supposed to be independent. However, for machine translation, there exist correlations between the target tokens, e.g., the syntax and grammar. In this concern, Sutskever et al. [262] proposed to use a shifted-right prediction scheme that an LSTM produces the target sentence after reading the input sequence.

2.1.1 RNN Encoder-Decoder

One of the difficulties of applying sequence-to-sequence models to machine translation is the alignment between the input and output sentence. For gauging the influence from both the input and preceding output tokens at each prediction, an additional RNN is placed parallel to the RNN of variable-length input, whose recurrence is activated by both the foregoing output and the last state of the RNN of input [47]. The RNN that sequentially reads the input is referred to as the Encoder, whilst the RNN that iterates over the output is referred to as the Decoder.

Despite being an effective framework, the RNN encoder is less efficient in aligning the semantics of output to those of input tokens, when compared with the additive attention [11], scilicet the first prototype of query-key-value attention [277], which computes the similarity between a target token (query) and a source token (key) as the normalised sum of corresponding hidden representations. The similarity score, i.e., the attention signal, is then used as the weight of the source token when summing up all the RNN encoder hidden states (values). Prediction is based on the updated RNN decoder hidden state activated by signals from both the previous hidden state and the aggregated RNN encoder.

The structure of the RNN Encoder-Decoder model is depicted in Figure 2.1, where the recurrent attention connects two RNNs. The attention mechanism is an aggregation over all the encoder RNN's hidden states weighted by similarity scores which usually takes the form of adding two corresponding hidden representations, as expressed as follows:

$$c_t = \sum_{n=1}^{N} \alpha_{tn} h_n \tag{2.1.1}$$

$$\alpha_{tn} = \frac{\exp(e_{tn})}{\sum_{n=1}^{N} \exp(e_{tn})} \tag{2.1.2}$$

$$e_{tn} = v^\top \tanh(W_s s_{t-1} + W_h h_n) \tag{2.1.3}$$

where $v \in \mathbb{R}^{d_a \times 1}$ is a weight vector, $s_{t-1} \in \mathbb{R}^{d_s \times 1}$ and $h_n \in \mathbb{R}^{d_h \times 1}$. W_s is a weight

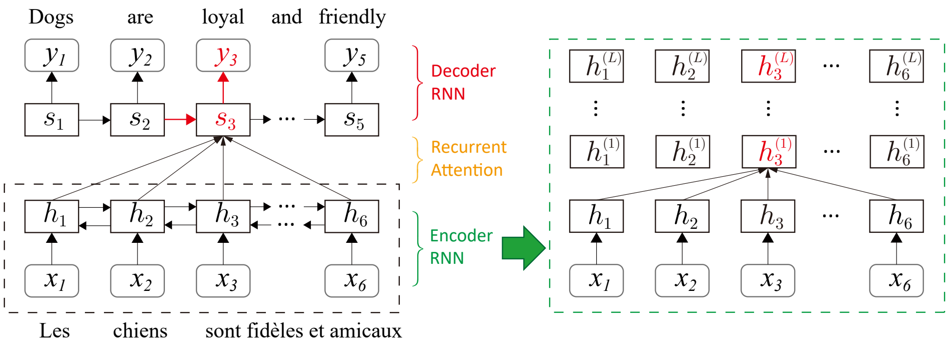


Figure 2.1: The RNN Encoder-Decoder architecture and the Self-Attention.

matrix of size $\mathbb{R}^{d_a \times d_s}$ and $W_h \in \mathbb{R}^{d_a \times d_h}$.

Intuitively, the attention mechanism is a soft alignment between the analysed vector and all the context vectors. Variants of this architecture often use different terminology to describe fundamentally similar ideas. For instance, Transformers [277] used dot-product attention [164] to calculate the similarity score as well as reduce the computational complexity. Rush et al. 2015 employed a weighted dot-product for alignment instead of an alignment MLP, and expand the decoder context from a single word to a context window. In the computationally less expensive multiplicative variant [164], Eq. 2.1.3 is expressed as

$$e_{tn} = (W_s s_{t-1})^\top W_h h_n \tag{2.1.4}$$

To this end, Galassi et al. 2021 summarised the nomenclatures and usages of various attention architectures.

2.1.2 Self-Attention

All the aforementioned attentions work under the scenario of machine translation, where the annotation is a sequence of tokens. It is also possible to apply the attention to a single sequence for the alignment between each word and other tokens in the sequence [230], as opposed to a single LSTM for the modelling of sequential dependence between words [30]. The right-hand side of Figure 2.1 shows several layers of Self-Attention. The upper layer representation is computed from inner alignment and weighted-sum of the lower layer, i.e., $c_t = h_t^{(1)}$ and $s_{t-1} = h_t$ in Eq. 2.1.1- 2.1.3.

If we confine the alignment between each word and every token to the alignment between these and a parameterised vector, Eq. 2.1.3 degenerates to

$$e_{tn} = v^\top \tanh(W_h h_n + b_h), \tag{2.1.5}$$

where b_h is analogous to a bias and v weighs the importance of each dimension. The attention above has demonstrated success in a spectrum of tasks replacing the pooling function to aggregate the final layer of Recurrent Neural Networks (RNNs). A notable work is the hierarchical attention networks (HAN) for document classification [305] where the Self-Attention is employed to produce a fixed-sized vector that represents sentences or documents respectively. Baziotis et al. 2017 subsequently applied Self-Attention to Twitter sentiment classification and achieved superior performance. Other works leverage Self-Attention to characterize the dependence between questions and answers in QA [160], to align passages and questions in Machine Reading Comprehension [240] and to attend to image patches in Image Caption Generation [301].

2.1.3 Memory Network

The neural attention can capture salient features among a collection of vectors, which naturally imitates the human brain behavior [8]. Unlike the LSTM's cell state that encodes relationship through recurrence, the neural attention allows the upper layer to directly attend to past hidden states, thus circumventing the LSTM's cell state bottleneck. This is analogous to a highway bypass the redundant connections or building blocks [90, 277, 278].

Despite the flexibility of skip-connections, the single-layered attention is inadequate for modelling the multiple hops over the long-term memory in comparison with LSTMs [259], let alone the temporal order. If we keep the LSTM for gauging the transitive dependencies, the recurrence will cost most of the computation since gradients propagating through RNN states must be calculated sequentially and cannot be fully parallelised [220]. To mitigate the RNN performance bottleneck, Sukhbaatar et al. 2015 developed the Memory Network that is fully composed of attention. In their model, as displayed in the right-hand side of Figure 2.1, the Encoder RNN is substituted with a stack of L Self-Attention layers. Hence, the RNNsearch [11] of multi-hop dependencies can be realised by bottom-up paths.

Inspired by this work, many sequence-to-sequence structures have replaced the recurrent block with layers of Self-Attentions, such as the Convolutional Sequence to Sequence model (ConvS2S) [72], the Gated Graph Sequence Neural Networks [149], and the Long Short-Term Memory-Networks [44]. In terms of text understanding tasks, the need for parallelisation in self-supervised learning on large unannotated corpora spurred interests in auto-regressive language models [271]. Narayan et al. 2018 introduced Memory Networks for Text Summarization to handle extremely-long dependencies. Similarly, the QA system developed by Miller et al. 2016 employed the key-valued memory network to index items from an external knowledge base. Similar memory network is also exploited in [161, 162].

2.1.4 Transformer Encoder-Decoder

Memory networks circumvent the computation for long-range recurrence of RNN states, at the cost of negligible performance impairment thanks to the parallelised gradient descent on self-attentions. Despite the intriguing properties, memory networks are less effective in jointly modelling the recurrent dependencies between labels and the alignment between the source sequence and target sequence, such as machine translation [297], due to the absence of the decoder structure [277], and caused by the limited expressiveness of stacking self-attentions when compared with stacked LSTMs [329].

Encoder

A natural way to introduce position relationships to pure attention is to use the positional encoding [245]. The word representations are added with positional encodings, so that the attention network could learn which position to attend to. Formally, the input h_i to a self-attention layer is decomposed as

$$h_i = x_i + p_i, \tag{2.1.6}$$

where p_i is the positional encoding which Vaswani et al. 2017 choose to be a sinusoid function and BERT [60] chooses to be a trainable embedding, x_i is the fixed word embedding.

On top of this, Vasawani et al. 2017 expanded the self-attention to the multi-head self-attention layer to allow information flows from different subspaces. Let the matrix $H \in \mathbb{R}^{N \times d_H}$ represent N rows of hidden representations (i.e., $[h_1; h_2; \ldots; h_N]$), the self-attention of Eq. 2.1.5 is then expressed as:

$$\text{head}_i = \text{Attention}(H) = \text{softmax}\left(\frac{HW_Q(HW_K)^\top}{\sqrt{d_h}}\right) HW_V, \tag{2.1.7}$$

$$\text{MultiHead}(H) = [\text{head}_1, \text{head}_2, \ldots, \text{head}_h] W_O, \tag{2.1.8}$$

where $W_Q \in \mathbb{R}^{d_H \times d_K}$, $W_K \in \mathbb{R}^{d_H \times d_K}$, $W_V \in \mathbb{R}^{d_H \times d_V}$, $W_O \in \mathbb{R}^{hd_V \times d_V}$, and $\text{softmax}(\cdot)$ denotes the softmax on each row. Note that here we denote hidden representations by $h_n \in \mathbb{R}^{1 \times d_H}$. This is different from the vertical vector representations in Eq. 2.1.1- 2.1.3. Most literature [45, 224, 323] refer to HW_Q, HW_K

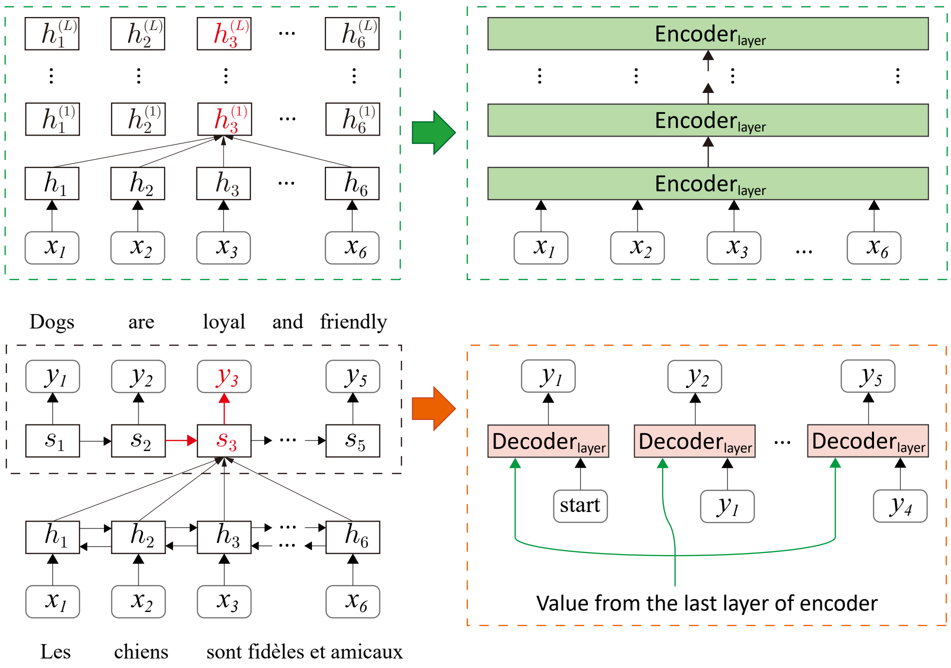


Figure 2.2: The Transformer Encoder-Decoder structure.

and HW_V as queries, keys and values. We will use this nomenclature in future discussions.

The recurrent attention of RNN Encoder-Decoder sidesteps redundant connections between each decoder state and historical encoder states, similarly, such gradient shortcuts could also exist across several layers when there are stacked self-attentions. In the Transformer Encoder-Decoder framework, as illustrated in Figure 2.2, each Encoder layer is integrated with a residual connection [90] placed in between its multi-head attention layer and that of the upper Encoder layer. This skip connection enables direct information flows to avoid gradient vanishing or explosion problems when the encoder goes deep. To further prevent exploding gradients and stabilize learning towards convergence, layer normalization [10] is employed after the residual connection.

Apart from the residual network, each Encoder layer consists of an identical fully-connected feed-forward network which takes input from token embeddings at each token position. This network is set to discern the influence of a token across different layers.

The most desirable advantage of stacking the Transformer Encoder layers compared with deep LSTMs is mass parallelisation. The direct connection between positions allows for parallelisation, and the dot-product attention is more efficient than its additive counterpart yet leading to no performance drop credited to the non-linearity of the residual connection plus ReLU layers.

Decoder

Memory network is proposed for Question Answering where the prediction is a single label. For Language Modelling [294], stacking Transformer Encoders is prevalent as most correlations are reflected in tuned masking strategies [6, 211]. In the context of Machine Translation, however, there has to be a connection in the label side since the prediction is no longer auto-regressive and label-to-label connection is indispensable. In this regard, Vasawani et al. 2017 designed the Decoder network to consume the shifted-right output from the proceeding Decoder layer. As depicted in Figure 2.2, the decoder comprises T layers. Each layer accepts prediction from the previous layer as the input for query, and the output matrix from the last encoder layer as inputs for keys and values. The attention in a decoder layer is expressed as:

$$\text{Attention}(o_{t-1}, H, H) = \text{softmax}\left(\frac{o_{t-1}^{\top} W_Q (H W_K)^{\top}}{\sqrt{d_h}}\right) H W_V, \tag{2.1.9}$$

where o_{t-1} is the output from the previous decoder layer. H is the output coming from the last layer of the encoder.

2.1.5 Recent Advances in Transformers

The efficiency of Transformer gave rise to pre-trained language models in large magnitudes [60, 91, 132, 155, 218, 304], since ELMo [213] had refreshed the state of the art on a suite of NLP tasks using pre-trained deep BiLSTMs. In lieu of the reduction in complexity [127], Sparse Attention [308] is proposed where each query attends to a random set of keys. Performer [49] employs kernel function to convert attention matrices to kernels, which simplifies matrices multiplication.

Another driving force behind applications of Transformers to language models is the refinement in modelling token order or token positions. In this avenue, Shaw et al. 2018 designed relative position embedding to explicitly model the relative position in matrix multiplication. Yang et al. 2019 refined the relative position embedding by adding a bias to the query matrix, which emphasizes the query-to-key position. More recently, He et al. 2021 developed the disentangled attention, namely DeBERTa, for token-to-token and token-to-position relations. The DeBERTa model also leverages absolute position for MLM prediction, and is the first to outperform the human baseline in the SuperGLUE [285] benchmark.

Improving transformer expressiveness is also a notable series of efforts. Henry et al. 2020 addressed the saturation of softmax function in the case where the dot-product results are congruently large, and remedied it with layer normalization. The same layer normalization is applied in replacement of the softmax layer to the dot-product attention [228]. It was later argued that layer normalization is insufficient to circumvent the rank collapse [62]. Therefore, residual connections and Multi-Layer Perceptions (MLPs) are essential for sustaining the rank of H.

2.2 Topic Representation Learning

Topic representation learning aims to encode a sentence into a low-dimensional space. The idea of encoding text into topic representations dates back to topic models [23, 24, 117] where texts are represented as a distribution over the latent variables. Earlier work explored Bayesian models or Bayesian nonparametric models for the generative process of text observations [22, 84, 202, 267]. The merit of generative models on topic representation learning is three-fold. Firstly, generative models can handle the explosive amount of unlabelled data effectively and hence, fit the training instances without the loss of generalization capability [312], as evidenced by Bayesian nonparametric models. Secondly, the modelling of latent variables enables disentanglement and causal reasoning of latent factors [238]. Thus, it is opportune in such a framework to impose prior distributions or steer the generation by tweaking latent factors. Lastly, the latent representations, whether disentangled or not, can comply with various mathematical constraints, which allow similarity-based operations such as clustering or semantic search [167, 187, 290, 335].

According to the OpenAI taxomony [120] and recent surveys [26, 122], existing deep generative models are categorized into five strands of work. The first, **autoregressive models**, are considered as latent-variable-free models not involving any probabilistic generative process. Token generation is formulated as shifted-right prediction or Denoising Auto-Encoders, such as Masked Language Models (MLMs) [187, 287, 306, 333]. The second kin are **normalising flows** [61], which assumes observations to be generated from samples of a latent variable through a chain of invertible functions [124]. The third, **Generative Adversarial Nets (GANs)** [78], introduces a discriminator to discriminate between the real instances and the generated samples, which uses adversarial training alternating between updating the discriminator and the generator. The fourth family of models are **Energy Based Models (EBMs)** [133], where the optimised probability $p_\theta(x) = \exp(-E_\theta(x))/\int_x \exp(-E_\theta(x))$. EBMs customize the Energy Function $E_\theta(x)$ and optimize θ based on the derivative of the log likelihood expressed as $\partial \log p_\theta(x)/\partial\theta = \mathbb{E}_{p_\theta(x)}\left[\partial E_\theta(x)/\partial\theta\right] - \partial E_\theta(x)/\partial\theta$. *Score-based model* [255] can be seen as an extension that surrogates the optimisation of $p_\theta(x)$ by optimizing a score function [153], thus circumventing the normalisation of $\exp(-E_\theta(x))$. The fifth, **Variational Auto Encoders (VAEs)** [123], stems from the posterior estimation in Bayesion nonparametrics, using parameterized inference network to maximize the Evidence Lower Bound (ELBo) of $p_\theta(x)$. *Diffusion models* can be viewed as an extension where the forward generative network contains multiple hops on the latent variables through Bayesian neural networks.

Of the most relevancy to this thesis is the VAE model. Like Energy Based Model, VAE hinges on a decomposition of $p_\theta(\mathbf{x})$, which is expressed as[1] [23, 25]:

$$\log p_\theta(\mathbf{x}) = \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{z},\mathbf{x}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right], \tag{2.2.1}$$

where the second RHS term is called the Evidence Lower BOund (ELBO). It is worth noting that Eq. 2.2.1 was originally discovered in Variational Inference [23, 117] during the decomposition of $\mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right]$[2], and was collated by Kingma et al. 2014 for generality, as will be introduced in § 2.2.2.

Eq. 2.2.1, which gives rise to this family of Probabilistic Graphical Models (PGMs), can be understood from two perspectives. For the perspective of $\log p_\theta(\mathbf{x})$, it is deemed as Maximum Likelihood Estimation (MLE) of θ [123]. For the perspective of $\mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right]$[3], one can regard it as Bayesian Inference of the posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ (whose point estimation is maximum a posteriori, scilicet MAP) [282].

In the presence of the latent variables, the marginal likelihood over $\mathbf{z}$, that is $\int_{\mathbf{z}} p_\theta(\mathbf{x},\mathbf{z})$, is typically intractable, leaving either the optimisation of $\log p_\theta(\mathbf{x})$ or $\mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right]$ an ill-posed problem [117, 282]. Thankfully, due to the positivity of $\log p_\theta(\mathbf{x})$ and $\mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right]$, both can resort to the optimisation of ELBO as an approximation, regardless of the trade-offs between the two objectives.

It is also possible that the generative process, AKA $p_\theta(\mathbf{x}|\mathbf{z})$ and $p_\theta(\mathbf{z})$, contains no parameters, and such an assumption favors some situations: if the data are synthetically generated, or the real-world data is generated by known distributions (e.g., tossing a dice) a priori, the latent variable model with parameterized genera-

[1] In LDA, the generative process is nonparametric. Therefore $p(x)$ does not include θ.

[2] Note that the decomposition can also be applied to the case $q_\phi(\mathbf{z})$, i.e., where we do not condition on $\mathbf{x}$ [123].

[3] In Variational Inference, the variational distribution is denoted as $q(\mathbf{z}|\gamma,\phi)$. It is important to note that the variational distribution is actually a conditional distribution [23], i.e., γ and ϕ are functions of $\mathbf{x}$ after the optimization has been conducted.

tive process will be over parameterized. From this fact, we dichotomize the PGMs into two categories – Bayesian models and Bayesian neural networks.

In the context of Bayesian models, the training objective can be understood as Bayesian inference. However, the object of finding $p_\theta(\mathbf{z}|\mathbf{x})$ is not restricted to approximation via a variational distribution, but is allowed to estimate from samples of the de facto posterior. In this sense, Gibbs Sampling [73, 74] can be employed to detour the intractability. In particular, Griffiths and Steyvers 2004 applied Collapsed Gibbs Sampling [152] to obtain samples from $p_\theta(\mathbf{z}|\mathbf{x})$.

The rest of this section will review the Variational Auto-Encoders in details, and give a brief introduction of posterior estimation methods for Bayesian models.

2.2.1 VAE for Topic Representation Learning

Let the Evidence Lower Bound be the target for optimization, the RHS term of Eq. 2.2.1 is rewritten as

$$\begin{aligned}
&\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_{\theta,\eta}(\mathbf{z},\mathbf{x}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right] && (2.2.2)\\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z}) + \log p_\eta(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right] && (2.2.3)\\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathbf{x}|\mathbf{z})\Big] - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log q_\phi(\mathbf{z}|\mathbf{x}) - \log p_\eta(\mathbf{z})\right] && (2.2.4)\\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})||p_\eta(\mathbf{z})\right], && (2.2.5)
\end{aligned}$$

where ϕ, θ and η are free parameters to be trained. Note that we consider η to be exclusive to $p_\eta(\mathbf{z})$ and $p_{\theta,\eta}(\mathbf{z},\mathbf{x}) = p_\theta(\mathbf{x}|\mathbf{z})\ p_\eta(\mathbf{z})$. It is assumed that $\{x_n\}_{n=1}^N$ and $\{z_n\}_{n=1}^N$ are i.i.d. variables in the generative process. This assumption leads to the factorization of the variational distribution. Therefore, Eq. 2.2.5 is expressed as

$$\mathbb{E}_{q_\phi(z_n|x_n)}\left[\log p_\theta(x_n|z_n)\right] - \mathrm{KL}\left[q_\phi(z_n|x_n)||p_\eta(z_n)\right], \tag{2.2.6}$$

where x_n is the BOW representation of the n-th document. The LHS term of Eq. 2.2.6 is commonly interpreted as an AutoEncoder [51, 123, 125, 227]. The variational distribution and the generative prior are commonly customized as Gaussian distributions such that

$$\begin{aligned}
z_n \sim q_\phi(z_n|x_n) &= \mathcal{N}(z_n; f_\mu(x_n), f_\sigma(x_n)) && (2.2.7)\\
p_\eta(z_n) &= \mathcal{N}(z_n; \mathbf{0}, \mathbf{I}), && (2.2.8)
\end{aligned}$$

where $f_\mu(\cdot)$ and $f_\sigma(\cdot)$ are MLPs parameterized by the variational parameters ϕ.

Under the BOW assumption, tokens are independent one-hot embeddings.

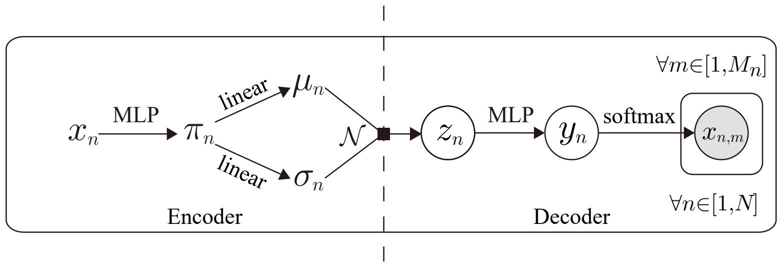


Figure 2.3: The plate diagram of VAE. Circled variables are random variables and those not circled are deterministic quantities. Shaded circles denote random observed quantities.

For the purpose of reconstructing the one-hot representation, the decoder is specialized [123] as

$$p_\theta(x_{n,m}|z_n) = \frac{\exp(y_n^{(x_{n,m})})}{\sum_{v=1}^{V}\exp(y_n^{(v)})}, \tag{2.2.9}$$

$$\text{where } y_n = \text{MLP}(z_n), \tag{2.2.10}$$

V is the vocabulary size, and $y_n^{(v)}$ is the v-th element of y_n. The BOW assumption and the specifications from Eq. 2.2.7 - 2.2.10 define the Neural Variational Document Model (NVDM) [179], a special case of VAE in the text understanding regime. Figure 2.3 depicts the structure of VAE, where the decoder is instantiated to softmax MLP, and the encoder is a multivariate Gaussian with a diagonal covariance.

Reparameterization Trick

VAE relies on Stochastic Gradient Descent (SGD) to learn the parameters. However, the Monte Carlo estimate of the expectation term in Eq. 2.2.6 requires sampling. If we sample $z_n^{(s)} \sim \mathcal{N}(z_n; f_\mu(x_n), f_\sigma(x_n))$ directly, the derivatives w.r.t. the parameters (i.e., $\nabla_\phi \Sigma_{s=1}^{S} z_n^{(s)}/S$) would exhibit very high variance [123, 201], making ϕ unable to converge to a local optimal. To circumvent this, a reparameterization trick is developed:

$$z_n^{(s)} = f_\mu(x_n) + f_\sigma(x_n) \odot \epsilon^{(s)}, \tag{2.2.11}$$

$$\text{where } \epsilon^{(s)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2.2.12}$$

Here, $\odot$ denotes the element-wise product and $z_n^{(s)}$ denotes the s-th sample from $\mathcal{N}(z_n; f_\mu(x_n), f_\sigma(x_n))$.

Computing $-\mathrm{KL}[q_\phi(z_n|x_n)||p_\eta(z_n)]$

Since both $q_\phi(z_n|x_n)$ and $p_\eta(z_n)$ are Gaussian distributions, the KL-divergence term can be analytically computed as

$$-\mathrm{KL}[q_\phi(z_n|x_n)||p_\eta(z_n)] = \frac{1}{2}\sum_{d=1}^{D}\left(1 + \log(\sigma_n[d]^2) - \mu_n[d]^2 - \sigma_n[d]^2\right), \tag{2.2.13}$$

$$\text{where } \mu_n = f_\mu(x_n) \text{ and } \sigma_n = f_\sigma(x_n), \tag{2.2.14}$$

D denotes the dimensionality of μ_n, and $\sigma_n[d]$ is the d-th element of σ_n.

To this end, the ELBO for SGD to optimize w.r.t. ϕ and θ is

$$\left[\frac{1}{S}\sum_{s=1}^{S}\sum_{m=1}^{M_n}\log p_\theta(x_{n,m}|z_n^{(s)})\right] + \frac{1}{2}\sum_{d=1}^{D}\left(1 + \log(\sigma_n[d]^2) - \mu_n[d]^2 - \sigma_n[d]^2\right), \tag{2.2.15}$$

where $\mu_n = f_\mu(x_n)$, $\sigma_n = f_\sigma(x_n)$, $z_n^{(s)} = \mu_n + \sigma_n \odot \epsilon^{(s)}$ and $\epsilon^{(s)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

2.2.2 Bayesian Models for Topic Modelling

According to the taxonomy at the head of this section, the generative model will be further dichotomized into two categories – Bayesian models and Bayesian neural networks, if it follows Eq. 2.2.1 to decompose the log-likelihood. Bayesian models, especially Bayesian nonparametric models circumvent the parametric redundancy, and their convergence to true labels is backed by theoretical developments [24]. For large-scale topic modelling of text, the most widely adopted method is the Latent Dirichlet Allocation (LDA) [23].

Latent Dirichlet Allocation

The trait of a Bayesian model is mainly described by its generative process:

1. For each topic $k \in \{1, 2, \ldots, K\}$
   - Draw a $V-1$-dimensional simplex $\beta_k \sim \text{Dirichlet}(\eta)$
2. For each document $n \in \{1, 2, \ldots, N\}$
   - Draw a $K-1$-dimensional simplex $\theta_n \sim \text{Dirichlet}(\alpha)$
   - For each token $m \in \{1, 2, \ldots, M_n\}$
     * Draw a topic $z_{n,m}|\theta_n \sim \text{Discrete}(\theta_n)$
     * Draw a word (piece) $w_{n,m}|\beta_{z_{n,m}} \sim \text{Discrete}(\beta_{z_{n,m}})$

The beauty of the Latent Dirichlet Allocation is 2-fold. The simplex in the first stage of the generative process, β_k, is a distribution over the vocabulary. The symmetric Dirichlet prior induces the simplex to be highly concentrated on a few of the values, which makes the topics distinct and easy to interpret, as well as the topical distributions of each document. Secondly, the Dirichlet distribution is conjugate to the multinomial distribution, which will facilitate the inference or estimation by confining the posterior to another Dirichlet distribution.

As mentioned in the taxonomy, the learning of latent variables is often understood as Bayesian inference of $p(\mathbf{z}|\mathbf{x})$ in Eq. 2.2.1, where two approaches – Collapsed Gibbs Sampling and Variational Inference are widely adopted. The variational inference is of the most relevance to this thesis. The variational inference for LDA typically makes the mean-field assumption that the variational family factorizes as

$$q(\boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:N}, \mathbf{z}_{1:N}|\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}) = \prod\nolimits_{k=1}^{K} q(\beta_k|\lambda_k) \prod\nolimits_{n=1}^{N} q(\theta_n|\gamma_n) \prod\nolimits_{m=1}^{M_n} q(z_{n,m}|\phi_{n,m}), \tag{2.2.16}$$

where $q(\beta_k|\lambda_k)$ and $q(\theta_n|\gamma_n)$ are Dirichlet distributions, since the posteriors of β_k and θ_n are Dirichlet distributions according to the Dirichlet-multinomial conjugate and Bayes Ball rules. $q(z_{n,m}|\phi_{n,m})$ is a discrete distribution. The variational parameters can be solved by a fixed-point iteration method which firstly iterates over n maximizing ELBO w.r.t. $\boldsymbol{\phi}_n$ and θ_n, and secondly iterates over k maximizing ELBO w.r.t. β_k.

Inspirations for VAE-based Topic Modelling

VAE also maximises ELBO but specifies $p_\eta(z_n) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, which is suboptimal compared with the Dirichlet distribution. However, choosing the Dirichlet distribution as the prior $p_\eta(z_n)$ raises two challenges: If we set $q(\cdot)$ to a Dirichlet family as well, it will be difficult to apply the reparameterization trick. While if we choose Gaussian distributions as the variational family $q(\cdot)$, the calculation of the KL-divergence between $p(\cdot)$ and $q(\cdot)$ would be more problematic [258]. ProdLDA [258] sidesteps these challenges by applying the logistic normal distribution to $p_\eta(z_n)$, which can approximate a Dirichlet distribution when its parameters are resolved by a closed form using a Laplace approximation [258]. Therefore, both $p(\cdot)$ and $q(\cdot)$ are set to $\mathcal{LN}(\cdot)$, and both the RT and computation of $\mathrm{KL}[q_\phi(z_n|x_n)||p_\eta(z_n)]$ become viable.

2.2.3 Recent Advances in VAE

AE and β-VAE

VAE optimises Eq. 2.2.6. However, it is also possible to learn a latent-feature discriminative model [125] by stressing the optimisation of $\mathbb{E}_{q_\phi(z_n|x_n)}[\log p_\theta(x_n|z_n)]$ and relaxing the constraint of $\mathrm{KL}[q_\phi(z_n|x_n)||p_\eta(z_n)]$. On the other hand, if we specify $q_\phi(z_n|x_n)$ to be a Dirac delta distribution (or its approximate equivalent) where all the probability mass is placed at $z_n = f_\phi(x_n)$, the ELBO will reduce to an AutoEncoder with the deterministic function $f_\phi(x_n)$ as the Encoder and $\log p_\theta(x_n|z_n)$ as the Decoder, and $\log p_\eta(f_\phi(x_n))$ will be a regularizer in which $p_\eta(\cdot)$ is often realised as $\mathcal{N}(z_n, \boldsymbol{\delta})$ quantifying the complexity of the encoder function.

In contrast to the AutoEncoder, β-VAE [94] hinges on the choice of $p_\eta(z_n)$, which serves as the prior to presumably induce the disentanglement (§ 2.3) of the latent variable. Successful disentanglement requires each component to correspond to an interpretable distribution or tangible factor, which is reflected in the KL-divergence term. Therefore, the loss objective is modified into

$$\mathbb{E}_{q_\phi(z_n|x_n)}\left[\log p_\theta(x_n|z_n)\right] - \beta\ \mathrm{KL}[q_\phi(z_n|x_n)||p_\eta(z_n)]\,, \tag{2.2.17}$$

where β is a hyperparameter controlling the strength of the correspondence. Higgins et al. 2017 showed that $\beta > 1$ is a typical value to achieve good disentanglement, indicating that unsupervised disentanglement heavily relies on the prior distribution.

VRNN

BOW representations of sentences omit the chronological order of tokens. Thus the word meanings depend solely on co-occurrences, which inhibits the full potential of representation learning. Conversely, it has been demonstrated that RNNs are more capable of language modelling and hence learn more compact representations compared with the vanilla Back Propagation (BP) network [18, 82, 180]. In the same vein, latent codes of tokens (i.e., $z_{n,m}$) shall have interrelations and such dependencies are essential for the modelling of word-level or sentence-level semantics.

To this end, Chung et al. 2015 proposed a recurrent version of VAE, called Variational Recurrent Neural network (VRNN), to explicitly model the dependencies between latent random variables across subsequent timesteps. It is assumed that both the generative network $p_\theta(x_t|\boldsymbol{z}_{\leq t}, \boldsymbol{x}_{<t})$ and the inference network $q_\phi(z_t|\boldsymbol{z}_{<t}, \boldsymbol{x}_{\leq t})$ are reliant on hidden states of an RNN. The resulting ELBO is expressed as

$$\mathbb{E}_{q_\phi(\boldsymbol{z}_{\leq T}|\boldsymbol{x}_{\leq T})}\left[\sum_{t=1}^{T}\left(\log p_\theta(x_t|\boldsymbol{z}_{\leq t}, \boldsymbol{x}_{<t}) - \mathrm{KL}(q_\phi(z_t|\boldsymbol{z}_{<t}, \boldsymbol{x}_{\leq t})||p_\theta(x_t|\boldsymbol{z}_{<t}, \boldsymbol{x}_{<t}))\right)\right]. \tag{2.2.18}$$

VAE with Normalizing Flows

The optimisation of ELBO requires the KL-divergence term of Eq. 2.2.6 to reach 0, which is hard because of the limited choices of approximating families (i.e., $q_\phi(z_n|x_n)$) or the assumed priors. An ideal variational family would be a flexible one that could contain the posterior distribution while minimising the KL divergence w.r.t. the prior. Therefore, Rezende and Mohamed 2015 introduced normalizing flows [263, 264] as the variational family. The normalizing flow transforms a base variational distribution, e.g., a sphere Gaussian distribution, to a multi-modal distribution through a sequence of invertible maps expressed as

$$\mathbf{z}_K = f_K \circ f_{K-1} \circ \cdots \circ f_1(\mathbf{z}) \tag{2.2.19}$$

$$\ln q_K(\mathbf{z}_K) = \ln q_0(\mathbf{z}) - \sum_{k=1}^{K} \ln |1 + \mathbf{u}_k^\top \phi_k(\mathbf{z}_{k-1})|, \tag{2.2.20}$$

where $\phi_k(\mathbf{z}) = h'(\mathbf{w}^\top \mathbf{z} + b)\mathbf{w}$ and $h'(\cdot)$ is the derivative of a smooth element-wise non-linearity. Then the ELBO expression in Eq. 2.2.2 can be rewritten as

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_{\theta,\eta}(\mathbf{z},\mathbf{x}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right] \tag{2.2.21}$$

$$= \mathbb{E}_{q_0(\mathbf{z}_0)}[\log p_{\theta,\mu}(\mathbf{x},\mathbf{z}_K) - \ln q_K(\mathbf{z}_K)] \tag{2.2.22}$$

$$= \mathbb{E}_{q_0(\mathbf{z}_0)}[\log p_{\theta,\mu}(\mathbf{x},\mathbf{z}_K)] - \mathbb{E}_{q_0(\mathbf{z}_0)}[\ln q_0(\mathbf{z}_0)] + \mathbb{E}_{q_0(\mathbf{z}_0)}\left[\sum_{k=1}^{K} \ln |1 + \mathbf{u}_k^\top \phi_k(\mathbf{z}_{k-1})|\right]. \tag{2.2.23}$$

Henceforth, the prior can be included in the variational family along with flexibility, i.e., choices other than Unit Gaussian, and the variational distribution will have more expressiveness.

VAE with Pre-Trained Language Models

Enriching the expressiveness of the prior and the variational distribution has garnered considerable interest in VAE research. In the wake of pre-trained language models, it is feasible to graft LM components to the inference network and formulate the training as fine-tuning in a low-resource setting. A practical approach is OPTIMUS [140], which is a modification of β-VAE that the inference network (i.e., Encoder) is BERT and the generative network (i.e., Decoder) is GPT-2. Notably, they design the `Memory Scheme` and the `Embedding Scheme` to regularize the latent variable with the variational prior. The `Memory Scheme` is operated by firstly reparametrizing z_n and then appending the sample to the LM output as an extra token. On the other hand, the `Embedding Scheme` adds the sample of the variational distribution to the LM output as a positional embedding. A number of recent papers [167, 187, 287] also addressed the fine-tuning of LM under the VAE formula, whilst their reconstruction objective is predicting the masked token in the same way the original transformer was trained instead of training from scratch. Hence we categorize them into Denoising Auto-Encoders and will discuss them in § 2.4.3.

Aside from grafting LM components, there are approaches utilizing off-the-shelf structures or embeddings. TopicBERT [39] concatenates the NVDM embedding with the `[CLS]` embedding of BERT to classify news articles into topics. TBERT [207] reconstructs the BERT embeddings and shows the usefulness of topic representations in paraphrase ranking. Bianchi et al. [20] proposed Neural Topic Models with Language Model Pre-training (NTMLM). The model is an extension of BOW neural topic models which consumes the concatenation of SBERT [225] embedding and BOW representation to reconstruct the BOW representation. TC-CTM [190] predicts the BOW representation with a similar architecture but with an added fully-connected layer and softmax to produce a topic classification from the LM hidden representations. In contrast, VIBERT [167] gave up the BOW reconstruction and only predicted a sentence label from the variational distribution, as tailored for the low-resource fine-tuning scenario.

2.3 Disentangled Learning

Deep learning methods in NLP learn the hidden semantics of text, many of which attempt to capture the independent latent factor to steer the generation of text [103, 115, 143, 210]. The ability to distinguish factors of variation from uninformative ones is called disentanglement [19, 95, 172]. The idea of disentangling independent components from their mixtures dates back to the linear Independent Component Analysis [109], where multiple linearly mixed signals can be recovered to their original source signals given that the source signals are non-Gaussian. Nowadays, there is surging interest in non-linear ICA [101, 107]. The majority of the work employs VAE [123] to learn controllable factors [35, 43, 94], as illustrated in β-VAE (§ 2.2.3) where a scaling hyperparameter is placed to align the variational distribution to a controllable prior.

However, theoretical studies on identifying factors of variations show that unsupervised learning of disentanglement by optimising the marginal likelihood in a generative model is impossible [157–159]. On the other hand, inductive biases from additional auxiliary variables or contrastive samples are helpful for extracting the underlying latent variables from data [108]. To solve the identifiability problem, Khemakhem et al. [121] proposed a premise on the observed marginal density $p_\theta(\mathbf{x})$ to offer the identifiability guarantees, which is specified by $\forall(\theta,\theta'):\ p_\theta(\mathbf{x}) = p'_\theta(\mathbf{x}) \ \Rightarrow\ \theta = \theta'$. In other words, if any two different choices of model parameter lead to the same marginal density, then they are equal and thus the models have the same joint distributions $p_\theta(\mathbf{x}, \mathbf{z})$. Therefore, the situation that two solutions share the same marginal density (i.e., $p_\theta(\mathbf{x})$) whilst convertible up to a transformation and thus entangled will be prevented under this assumption.

More recently, Horan et al. [97] proposed to unleash the constraint on identifiability to a more general assumption – the assumption of *local isometry*, that any change in the latent variable is associated with a change in the observation. The local isometry suffices to find a disentangled representation even with classical methods such as FastICA. While the correspondence between the variable of variations and the observations induces disentanglement, the statistical correlations between observed factors of variations pose problems for generative models attempting to learn a disentangled representation. Träuble et al. [273] empirically studied these effects and investigated two approaches to resolve the correlations.

2.4 Semi-supervised Language Representation Learning

Encoding words or other component units of language into compact, exploitable representations has been the central theme of text understanding research. In general, it is feasible to use representations acquired from self-supervised [88, 181] or unsupervised learning [242] to codify the semantics for supervised learning, or even to pave the foundation for reinforcement learning, as evidenced by recently developed pre-trained language models such as PaLM and GPT-3.5 [34, 50, 199]. In this section, we first introduce word embedding and language modelling approaches that build foundation representations for supervised learning or fine-tuning. Then we proceed with literature reviews on Denoising Auto-Encoders.

2.4.1 Word Embedding

Word-level representation learning aims at encoding words through a lookup table into a low-dimensional space as vectors or densities. Normally, word representations are encoded from the collocations of words in the local context, and thus can presumably summarize the syntactic and semantic regularities by observing the co-occurrence. In this way, the dense representations of words will transfer the encoded statistics, usually reflected by word similarities, to boost the performance of downstream tasks. Successful applications include question answering [52], textual entailment [77], named entity recognition [131] and sentiment analysis [232], where the embedded words serve as input to models for downstream tasks. The idea of learning word representations by backpropagation (BP) neural networks was first explored in [234]. Later on Deerwester et al. [1990] addressed this problem by latent semantic indexing where word embeddings can be extracted by performing singular vector decomposition on word co-occurrence matrices. A notable approach is the Skip-Gram model, also known as *Word2Vec* [181], which inherits the idea of parametric vocabulary-to-vector mapping in the neural language model [18]. More concretely, their model optimizes the function $\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} log\, p(w_{n,c}|x_n)$, to maximize the log-likelihood of the context $\mathbf{w}_n$ given the n-th word x_n. The Skip-Gram was further modified to scale up to large amounts of data by replacing the softmax layer with hierarchical softmax or Negative Sampling (NEG) [182]. Pennington et al. [2014] pointed out that Skip-Gram only utilizes local context while ignoring the document-level word co-occurrence counts. Their proposed GloVe model integrates the matrix factorization by modelling local context likelihood as document-dependent.

Meanwhile, there is an emerging tendency towards applying density representations to discriminate the nuance of information among senses. For example, the Gaussian word embedding proposed in [280] represented words directly as Gaussian distributions. Barkan [2017] extended the Skip-Gram by placing a Gaussian prior on the parameterized word vectors. The parameters were learned via variational inference [117]. Their model is the first to formulate the parameter optimization problem as a posterior inference, which is typically used in probabilistic graphical models. Bražinskas et al. [2018] represented words as posterior densities conditioned on the pivot word and the associated context.

2.4.2 Pre-Trained Language Models

Language modelling is an effective approach to generating language representations, usually at the token level. Unlike the word embedding, the objective of a language model is to predict a joint distribution over a sequence of words $\{w_1, w_2, \ldots, w_T\}$: $p(w_1, w_2, \ldots, w_T) = \prod_{t=1}^{T} p(w_t|w_1, w_2, \ldots, w_{t-1})$, which is essentially a generative model and can be categorized into autoregressive models already defined in § 2.2. In the case of computing the joint distribution by predicting the masked tokens, as practised in pre-trained MLMs [60], they are generalised as Denoising Auto-Encoders. Though in this subsection we still refer to them as MLMs. In § 2.4.3, we will discuss a variant of Denoising Auto-Encoders and distinguish them from MLMs.

To compute the conditional probability of the autoregressive model, traditional approaches use non-parametric N-gram smoothing models [80]. Bengio et al. [2003] made the first attempt to utilise a vanilla Back Propagation (BP) network for the N-gram autoregressive prediction and vectorized the vocabulary via a look-up table. Mikolov et al. [2010] followed their work by employing a Recurrent Neural Network (RNN) and a Softmax function to predict the conditional probability. Most of the work was limited to training the context-free word representations (i.e., the look-up table) at this stage until Howard and Ruder [99] discovered that the hidden states of the top layer of pre-trained LSTMs could be directly employed for classification and performance could be further improved by adding an extra layer for fine-tuning. They refreshed the state-of-the-art in multiple domain-agnostic tasks. Finally, Peters et al. [213] proposed ELMO that exploited aggregated hidden states of deep BiLSTMs and unified the fine-tuning process. They showed that pre-trained BiLMs provide superior representations or network weights beneficial for downstream classification, whilst fine-tuning improves domain-specific performance at the cost of perplexity impairments. More importantly, by probing into different layers of the pre-trained model, they found that these layers encode semi-supervision signals at different abstraction levels (i.e., higher-level LSTMs capture semantics while lower-level states encode the syntax).

The success of ELMO inspired a broad range of semi-supervised frameworks that ameliorate LM pre-training and fine-tuning. A family of them employed advanced Sequence-to-Sequence models (e.g., Transformers), among which we have GPT/GPT-2 [218, 219], BERT [60], RoBERTa [155], XLNet [304], ALBERT [132], T5 [221], and so on [294]. In particular, DeBERTa [91] designed disentangled attention by itself and encoded the positional and word embeddings separately. From the data-format perspective, Wolf et al. [294] distinguish between autoregressive prediction and masked-token prediction in language models, suggesting that autoregressive LMs and MLMs are akin to sequence-to-sequence models, whilst they only differ in the format of input-output token sequences. In this sense, the sequence-to-sequence relationship modelled by autoregressive LMs are shifted-right token-level dependencies between the output and either the inputs or the preceding outputs. For autoregressive LMs, the output-output dependencies can be implicatures of output-input relationships. For MLMs, those relationships reside in the cooccurrence of non-special tokens and `[MASK]` tokens. It should be noted that dependencies among the output sequence are modelled by the Decoder, and in Transformers of Encoders only, what LMs can learn is limited to target-input sequence dependencies. This accounts for the phenomenon that GPTs [34] comprise Decoders as building blocks. On the other hand, Encoder-based LMs typically make consecutive predictions as the word prediction mechanism, or employ mask-tune [6, 137] to compensate for the output sequence dependency.

Recently, Large Language Models (LLMs) that have unified different tasks as free-form generation or multiple-choice selection [34, 270], have demonstrated remarkable success. The rationale is that massive training corpora encompass a mixture of implicit tasks that LLMs can learn in the process of learning to predict the next word. In particular, GPT-3 [34] features in-context learning to navigate agnostic tasks. The goal of in-context learning is to allow the model to do a completion given a prompt in a specific context formulated as $p_\theta(\mathsf{completion}|\mathsf{prompt}, \mathsf{context})$, where $\mathsf{context}$ is instantiated as free-form task descriptions and/or few-shot examples illustrating the task, and prompt is a direct instruction for the $\mathsf{completion}$. GPT-3 carries out in-context learning by prepending examples with task descriptions or few-shot examples in the free-from texts. They rephrase or reformat the training examples of each task to fulfill a single training objective which does not cater to any task in particular. The resulting model, i.e., GPT-3, shows few-shot (or zero- and one-shot) ability. In this avenue, instruction-tuning [291] has been developed for the benefit of zero-shot prediction on unseen tasks. The instruction-tuning pipeline comprises a manual template creation step which composes natural language instructions to describe the task for each dataset, and a fine-tuning step which tunes a pretrained language model with examples from each dataset formatted via a randomly sampled instruction template for that dataset. Similarly, the T0 model [236] develops an interface for prompt collection and provides each dataset with multiple prompt templates. Finally, Ouyang et al. [199] design Reinforcement Learning from Human Feedback (RLHF) to mitigate the hallucination, i.e., to reduce the untrue output, which features a 3-stage pipeline that (1) an 175B GPT-3 model is tuned using prompts sampled from a prompt dataset, in the same way as instruction-tuning; (2) a 6B GPT-3 model is turned into a reward model, which learns from ranked outputs annotated by labelers to calculate the reward signal for each output; (3) the reward model updates the 175B GPT-3-driven completion policy using Proximal Policy Optimization (PPO).

There are also attempts leveraging knowledge graphs [237] or event databases [135] for the integrity of commonsense reasoning [28, 214]. The training and tuning setup has been optimised [155]. New benchmarks [284, 285] and metrics [318] have been proposed for extensive evaluation.

2.4.3 Denoising Auto-Encoders

Deep pre-trained transformers encode tokens into tunable general-purpose representations, which improve the target-domain performance. This semi-supervised framework has been demonstrated successful by a number of specific models [60, 91, 155]. However, token-level representations are susceptible to being overparameterized and cumbersome as appeared in Neural Sequence Models discussed in § 2.2. For example, averaging the BERT output layer (known as BERT embedding) or fine-tuning the standalone `[CLS]` are inefficient in semantic search where the evaluation is more about sentence-level similarities [225]. Text clustering also requires a fixed-length sentence representation compatible with the clustering algorithms [299]. To mitigate these, a bottleneck representation was introduced to distil the holistic properties of a sentence. Such a representation is typically a condensation over intermediate outputs, which is analogous to a pooling operation. Since MLMs pertain to Denoising Auto-Encoders according to the § 2.2 taxonomy, and GPT-alike LMs are essentially autoregressive models, their training objectives are compatible so that the semi-supervised learning could be performed.

Specifically, Montero et al. [187] presents a sentence bottleneck autoencoder, called AutoBot, which clamps the encoder representation into a fixed-size latent code. The latent code is learnt from the reconstruction of the perturbed text for the benefit of dynamically pooling semantic information from the pre-trained model's hidden states. Yang et al. [306] also developed a sentence representation encoder, where the sentence representation functions as a trainable vector to prompt a conditional masked language model. In another work, SentenceMIM [156] followed the denoising auto-encoding strategy but did the training from scratch. The model produces sentence representations suitable for similarity-based clustering and QA. TS-DAE [287] is another sentence embedding model targeting unsupervised fine-tuning whose objective is to predict the masked tokens and bottleneck representation is acquired from the fine-tuning. Recent LLMs, e.g., cpt-code [195] and E5 [288], have combined Denoising Auto-Encoders with in-batch contrastive learning and contrastive fine-tuning to learn dense representations and utilised such representations to index semantically-related pairs in their semantic-searching modular, which supports relevant code search with a query in natural language.

2.5 Applications

This chapter introduced the relevant models in a taxonomy where each branch is self-contained and chronologically updated. While these strands of work progress in diverse directions, some share intersections regarding modules and objectives, which attracts attention to particular tasks. For example, aspect-based sentiment analysis [203] often requires the identification of aspects/topics and their polarities, so that the joint modelling of aspects topics and sentiments [150] shows an advantage. Opinion extraction requires the disentanglement of aspect and sentiment [326]. Topic representation learning and language modelling collaboratively play a vital role in these scenarios. Another notable task is Text Clustering [299] which requires clustering-friendly representations [266]. In such a scenario, instructive annotation is laborious and often requires expert knowledge. Therefore, the majority of approaches resort to the semi-supervised learning paradigm that pretrains distributed representations first and then fine-tunes an LM-based Denoising Auto-Encoder. We will use these tasks as test beds in the following chapters to evaluate the proposed models.

Chapter 3

A Neural Generative Model for Joint Learning Topics and Topic-Specific Word Embeddings

Chapter Abstract

This chapter introduces a generative model to explore the local and global context for joint learning topics and topic-specific word embeddings. We assume that global latent topics are shared across documents, a word is generated by a hidden semantic vector encoding its contextual semantic meaning, and its context words are generated conditional on both the hidden semantic vector and global latent topics. Topics are trained jointly with the word embeddings. The trained model maps words to topic-dependent embeddings, which naturally addresses word polysemy. We show experiments on word similarity evaluation and word sense disambiguation, demonstrating the model's effectiveness in word representation learning. Besides word embeddings, the model extracts more coherent topics than existing neural topic models or other models for joint learning of topics and word embeddings.

3.1 Introduction

Probabilistic topic models assume words are generated from latent topics which can be inferred from word co-occurrence patterns taking a document as global context. In recent years, various neural topic models have been proposed. Some of them are built on the Variational Auto-Encoder (VAE) [123] which utilizes deep neural networks to approximate the intractable posterior distribution of observed words given latent topics [29, 179, 258]. However, these models take the bag-of-words (BOWs) representation of a given document as the input to the VAE and aim to learn hidden topics that can be used to reconstruct the original document. They do not learn word embeddings concurrently.

Other topic modeling approaches explore the pre-trained word embeddings for the extraction of more semantically coherent topics since word embeddings capture syntactic and semantic regularities by encoding the local context of word co-occurrence patterns. For example, the topic-word generation process in the traditional topic models can be replaced by generating word embeddings given latent topics [55] or by a two-component mixture of a Dirichlet multinomial component and a word embedding component [196]. Alternatively, the information derived from word embeddings can be used to promote semantically-related words in the Polya Urn sampling process of topic models [139] or generate topic hierarchies [324]. However, all these models use pre-trained word embeddings and do not learn word embeddings jointly with topics.

While word embeddings could improve the topic modeling results, but conversely, the topic information could also benefit word embedding learning. Early word embedding learning methods [181] learn a mapping function to project a word to a single vector in an embedding space. Such one-to-one mapping cannot deal with word polysemy, as a word could have multiple meanings depending on its context. For example, the word '*patient*' has two possible meanings '*enduring trying circumstances with even temper*' and '*a person who requires medical care*'. When analyzing reviews about restaurants and health services, the semantic meaning of '*patient*' could be inferred depending on which topic it is associated with. One solution is to first extract topics using the standard Latent Dirichlet Allocation (LDA) model and then incorporate the topical information into word embedding learning by treating each topic as a pseudo-word [154].

Whereas the aforementioned approaches adopt a two-step process, by either using pre-trained word embeddings to improve the topic extraction results in topic modelling, or incorporating topics extracted using a standard topic model into word embedding learning, Shi et al. [248] developed a Skip-Gram-based model to jointly learn topics and word embeddings based on the Probabilistic Latent Semantic Analysis (PLSA), where each word is associated with two matrices rather than a vector to induce topic-dependent embeddings. This is a rather cumbersome setup. Foulds [69] used the Skip-Gram to imitate the probabilistic topic model that each word is represented as an importance vector over topics for context generation.

In this chapter, we design a neural generative model built on VAE, called the Joint Topic Word-embedding (JTW) model, for jointly learning topics and topic-specific word embeddings. More concretely, we introduce topics as tangible parameters that are shared across all the context windows. We assume that the pivot word is generated by the hidden semantics encoding the local context where it occurred. Then the hidden semantics is transformed to a topical distribution taking into account the global topics, and this enables the generation of context words. Our rationale is that the context words are generated by the hidden semantics of the pivot word together with a global topic matrix, which captures the notion that the word has multiple meanings that should be shared across the corpus. We are thus able to learn topics and generate topic-dependent word embeddings jointly. The results of our model also allow the visualization of word semantics because topics can be visualized via the top words and words can be encoded as distributions over the topics.[1] In particular, we make the following contributions:

- We propose a novel Joint Topic Word-embedding (JTW) model built on VAE, for jointly learning topics and topic-specific word embeddings;
- We perform extensive experiments and show that JTW outperforms other Skip-Grams or Bayesian alternatives in both word similarity evaluation and word sense disambiguation tasks, and can extract semantically more coherent topics from data;
- We also show that JTW can be easily integrated with existing deep contextualized word embedding learning model to further improve the performance of downstream tasks such as sentiment classification.

3.2 Related Work

Skip-Gram approaches for word embedding learning The Skip-Gram, also known as WORD2VEC [182], maximizes the probability of the context words $\mathbf{w}_n$ given a centroid word x_n. Pennington et al. [209] pointed out that Skip-Gram neglects the global word co-occurrence statistics. They thus formulated the Skip-

[1] The code is accessible via `http://github.com/somethingx02/topical_wordvec_models`.

Gram as a non-negative matrix factorization (NMF) with the cross-entropy loss switched to the least square error. Another NMF-based method was proposed by Xu et al. [300], in which the Euclidean distance was substituted with the Wasserstein distance. Jameel and Schockaert [111] rewrote the NMF objective as a cumulative product of normal distributions, in which each factor is multiplied by a von Mises-Fisher (vMF) distribution of context word vectors, to hopefully cluster the context words since the vMF density retains the cosine similarity.

Although the Skip-Gram-based methods attracted extensive attention, they were criticized for their inability to capture the polysemy [216]. A pioneered solution to this problem is the Multiple-Sense Skip-Gram (MSSG) model [194], where word vectors in a context are first averaged then clustered with other contexts to obtain a sense representation for the pivot word. In the same vein, Iacobacci and Navigli [110] leveraged sense tags annotated by BabelNet [193] to jointly learn word and sense representations in the Skip-Gram manner that the context words are parameterized via a shared look-up table and sent to a BiLSTM to match the pivot word vector.

There have also been Bayesian extensions of the Skip-Gram models for word embedding learning. Barkan [15] inherited the probabilistic generative line while extending the Skip-Gram by placing a Gaussian prior on the parameterized word vectors. The parameters were estimated via variational inference. In a similar vein, Rios et al. [229] proposed to generate words in bilingual parallel sentences by shared hidden semantics. They introduced a latent index variable to align the hidden semantics of a word in the source language to its equivalence in the target language. More recently, Bražinskas et al. [32] proposed the Bayesian Skip-Gram (BSG) model, in which each word type with its related word senses collapsed is associated with a 'prior' or static embedding and then, depending on the context, the representation of each word is updated by 'posterior' or dynamic embedding. Through Bayesian modelling, BSG is able to learn context-dependent word embeddings. It does not explicitly model topics, however. In our proposed JTW, global topics are shared among all documents and learned from data. Also, whereas BSG only models the generation of context words given a pivot word, JTW explicitly models the generation of both the pivot word and the context words with different generative routes.

Combining word embeddings with topic modeling Pre-trained word embeddings can be used to improve the topic modelling performance. For example, Das et al. [55] proposed the Gaussian LDA model, which, instead of generating discrete word tokens given latent topics, generates draws from a multivariate Gaussian of word embeddings. Nguyen et al. [196] also replaced the topic-word Dirichlet multinomial component in traditional topic models, but by a two-component mixture of a Dirichlet multinomial component and a word embedding component. Li et al. [139] proposed to modify the Polya Urn sampling process of the LDA model by promoting semantically-related words obtained from word embeddings. More recently, Zhao et al. [324] proposed to adapt a multi-layer Gamma Belief Network to generate topic hierarchies and also fine-grained interpretation of local topics, both of which are informed by word embeddings.

Instead of using word embeddings for topic modeling, Liu et al. [154] proposed the Topical Word Embedding model which incorporates the topical information derived from standard topic models into word embedding learning by treating each topic as a pseudo-word. Briakou et al. [33] followed this route and proposed a four-stage model in which topics were first extracted from a corpus by LDA and then the topic-based word embeddings were mapped to a shared space using anchor words which were retrieved from the WordNet.

There are also approaches proposed to learn topics and word embeddings built on Skip-Gram models jointly. Shi et al. [248] developed a Skip-Gram Topical word Embedding (STE) model built on PLSA where each word is associated with two matrices—one matrix used when the word is a pivot word and another used when the word is considered as a context word. Expectation Maximization (EM) is used to estimate model parameters. Foulds [69] proposed the Mixed-Membership Skip-Gram model (MMSG), which assumes a topic is drawn for each context and the word in the context is drawn from the log-bilinear model based on the topic embeddings. Foulds trained their model by alternating between Gibbs sampling and noise-contrastive estimation. MMSG only models the generation of context words, but not pivot words.

While our proposed JTW also resembles the similarity to the Skip-Gram model in that it predicts the context word given the pivot word, it is different from the aforementioned approaches in that it assumes global latent topics shared across all documents, and the generation of the pivot word and the context words follows different generative routes. Moreover, it is built on VAE and is trained using neural networks for more efficient parameter inference.

3.3 Joint Topic Word-embedding (JTW)

In this section, we describe our proposed Joint Topic Word-embedding (JTW) model built on VAE, as shown in Fig. 3.1. We first give an overview of JTW, then present each component of the model, followed by the training details.

Following the problem setup in the Skip-Gram model, we consider a pivot word x_n and its context window $\mathbf{w}_n = w_{n,1:C}$. We assume there are a total of N pivot word tokens and each context window contains C context words. However, as opposed to Skip-Gram, we do not compute the joint probability as a product chain of conditional probabilities of the context word given the pivot. Instead, in our model, context words are represented as BOWs for each context window by assuming the exchangeability of context words within the local context window.

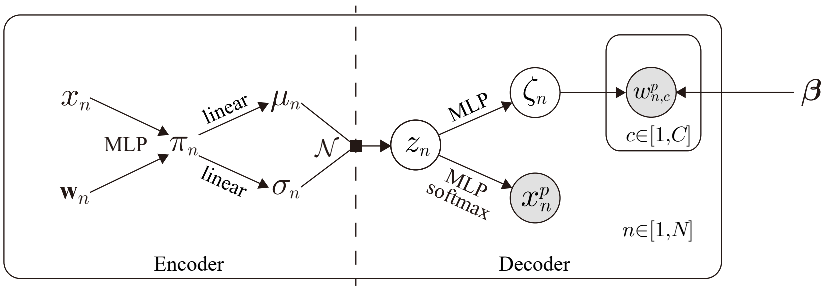


Figure 3.1: The Variational Auto-Encoder framework for the Joint Topic Word-embedding (JTW) model. Boxes are "plates" indicating replicates. Shaded circles represent the observed variables. $\boldsymbol{\beta}$ is a T $\times$ V matrix representing corpus-wide latent topics.

We hypothesize that the hidden semantic vector z_n of each word x_n induces a topical distribution that is combined with the global corpus-wide latent topics to generate context words. Topics are represented as a probability matrix where each row is a multinomial distribution measuring the importance of each word within a topic. The hidden semantics z_n of the pivot word x_n is transformed to a topical distribution ζ_n, which participates in the generation of context words. Our assumption is that each word embodies a finite set of meanings that can be interpreted as topics, thus each word representation can be transformed to a distribution over topics. Context words are generated by first selecting a topic and then being sampled according to the corresponding multinomial distribution. This enables a quick understanding of word semantics through the topical distribution and at the same time learning the latent topics from the corpus. The generative process is given below:

- For each word position $n \in \{1, 2, 3, \ldots, N\}$:
  - Draw hidden semantic representation $z_n \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  - Choose a pivot word $x_n \sim p(x_n|z_n)$
  - Transform z_n to ζ_n with a multi-layered perceptron: $\zeta_n = \text{MLP}(z_n)$

- For each context word position $c \in \{1, 2, 3, \ldots, C\}$:
  * Choose a topic indicator $t_{n,c} \sim \text{Categorical}(\zeta_n)$
  * Choose a context word $w_{n,c} \sim p(w_{n,c}|\beta_{t_{n,c}})$

Here, all the distributions are functions approximated by neural networks, e.g., $p(x_n|z_n) \propto \exp(\boldsymbol{M}_x z_n + \boldsymbol{b}_x)$, which will be discussed in more details in the Decoder section, $t_{n,c}$ indexes a row $\beta_{t_{n,c}}$ in the topic matrix. We could implicitly marginalise out the topic indicators, in which case the probability of a word would be written as $w_{n,c}|\zeta_n, \boldsymbol{\beta} \sim \text{Categorical}(\sigma(\boldsymbol{\beta}^{\mathrm{T}}\zeta_n))$, where $\sigma(\cdot)$ denotes the softmax function. The prior distribution for z_n is a multivariate Gaussian distribution with the mean $\mathbf{0}$ and covariance $\boldsymbol{I}$, of which the posterior indicates the hidden semantics of the pivot word when conditioned on $\{x_n, \mathbf{w}_n\}$.

Although both JTW and BSG assume that a word can have multiple senses and use a latent embedding z to represent the hidden semantic meaning of each pivot word, there are some key differences in their generative processes. JTW first draws a latent embedding z from a standard Gaussian prior which is deterministically transformed into topic distributions and a distribution over pivot words. The pivot word is conditionally independent of its context given the latent embedding. At the same time, each context word is assigned a latent topic, drawn from a shared topic distribution which leverages the global topic information, and then drawn independently of one another. In BSG the latent embedding z is also drawn from a Gaussian prior but the context words are generated directly from the latent embedding z, as opposed to via a mixture model as in JTW. Therefore, JTW is able to group semantically-similar words into topics, which is not the case in BSG.

Given the observed variables $\{x_{1:N}, \mathbf{w}_{1:N}\}$, the objective of the model is to infer the posterior $p(\mathbf{z}|\mathbf{x}, \mathbf{w})$. This is achieved by the VAE framework. As illustrated in Figure 3.1, the JTW model is composed of an encoder and a decoder, each of which is constructed by neural networks. The family of distributions to approximate the posterior is Gaussian, in which μ_n and σ_n are optimized. As in VAE, we optimize μ_n and σ_n through the training of parameters in neural networks (e.g., we optimize $\boldsymbol{M}_\pi$ in $\mu_n = \boldsymbol{M}_\pi^{\mathrm{T}}\pi_n + \boldsymbol{b}_\pi$ instead of updating μ_n directly).

3.3.1 ELBO

The VAE naturally simulates the variational inference [117], where a family of parameterized distributions $q_\phi(z_n|x_n, \mathbf{w}_n)$ are optimized to approximate the intractable true posterior $p_\theta(z_n|x_n, \mathbf{w}_n)$. This is achieved by minimizing the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior for each data point:

$$
\begin{aligned}
&\mathrm{KL}(q_\phi(z_n|x_n, \mathbf{w}_n)||p_\theta(z_n|x_n, \mathbf{w}_n)) \\
&= \log p_\theta(x_n, \mathbf{w}_n) - \mathbb{E}_{q_\phi}[\log p_\theta(z_n, x_n, \mathbf{w}_n) - \log q_\phi(z_n|x_n, \mathbf{w}_n)],
\end{aligned} \tag{3.3.1}
$$

where the expectation term is called the Evidence Lower Bound (ELBO), denoted as $\mathcal{L}(\theta, \phi; x_n, \mathbf{w}_n)$. VAE optimizes ELBO to presumably minimize the KL-divergence. The ELBO is further derived as

$$
\begin{aligned}
&\mathcal{L}(\theta, \phi; x_n, \mathbf{w}_n) \\
&= \mathbb{E}_{q_\phi(z_n|x_n, \mathbf{w}_n)}[\log p_\theta(x_n, \mathbf{w}_n|z_n)] - \mathrm{KL}(q_\phi(z_n|x_n, \mathbf{w}_n)||p(z_n)).
\end{aligned} \tag{3.3.2}
$$

The first term on the left-hand side of Equation 3.3.2, which is an expectation with respect to $q_\phi(z_n|x_n, \mathbf{w}_n)$, can be estimated by sampling due to its intractability. That is:

$$
\mathbb{E}_{q_\phi(z_n|x_n, \mathbf{w}_n)}[\log p_\theta(x_n, \mathbf{w}_n|z_n)] \approx \frac{1}{S}\sum_{s=1}^{S} \log p_\theta(x_n, \mathbf{w}_n|z_n^{(s)}), \tag{3.3.3}
$$

where $z_n^{(s)} \sim q_\phi(z_n|x_n, \mathbf{w}_n)$. Here we use $z_n^{(s)}$ to represent the samples since the sampled distribution is related to x_n.

3.3.2 Encoder

The Encoder corresponds to $q_\phi(z_n|x_n, \mathbf{w}_n)$ in Equation 3.3.3. Recall that the variational family for approximating the true posterior is Gaussian Distribution parameterized by $\{\mu_n, \sigma_n\}$. As such, the encoder is essentially a set of neural functions mapping from observations to Gaussian parameters $\{\mu_n, \sigma_n\}$. The neural functions are defined as: $\pi_n = \mathrm{MLP}(x_n, \mathbf{w}_n)$, $\mu_n = \boldsymbol{M}_\mu^{\mathrm{T}}\pi_n + \boldsymbol{b}_\pi$, $\sigma_n = \boldsymbol{M}_\sigma^{\mathrm{T}}\pi_n + \boldsymbol{b}_\sigma$, where the MLP denotes the multi-layered perceptron and the context window $\mathbf{w}_n$ is represented as a BOW that is a V-dimentional vector. The encoder outputs Gaussian parameters $\{\mu_n, \sigma_n\}$, which constitutes the variational distribution $q_\phi(z_n|x_n, \mathbf{w}_n)$. In order to differentiate $q_\phi(z_n|x_n, \mathbf{w}_n)$ with respect to ϕ, we apply the reparameterization trick [123] by using the following transformation:

$$
\begin{aligned}
z_n^{(s)} &= \mu_n + \sigma_n \odot \epsilon_n^{(s)} \\
\epsilon_n^{(s)} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}).
\end{aligned} \tag{3.3.4}
$$

3.3.3 Decoder

The Decoder corresponds to $p_\theta(x_n, \mathbf{w}_n|z_n^{(s)})$ in Equation 3.3.3. It is a neural function that maps the sample $z_n^{(s)}$ to the distribution $p_\theta(x_n^p, \mathbf{w}_n^p|z_n^{(s)})$ with random variables instantiated by x_n and $\mathbf{w}_n$. More concretely, we define two neural functions to generate the pivot word and the context words separately. Both the functions involve an MLP, while the context words are generated independently from each other by the topic mixture weighted by the hidden topic distributions. The neural functions are expressed as:

$$p(x_n^p|z_n^{(s)}) \propto \exp(\boldsymbol{M}_x z_n^{(s)} + \boldsymbol{b}_x) \tag{3.3.5}$$

$$\zeta_n^{(s)} = \mathrm{MLP}(z_n^{(s)}) \tag{3.3.6}$$

$$p(w_{n,c}^p|\zeta_n^{(s)}) \propto \exp(\boldsymbol{\beta}^{\mathrm{T}}\zeta_n^{(s)} + \boldsymbol{b}_w) \tag{3.3.7}$$

In this case, the MLP for the pivot word is specified as a fully-connected layer. Recall that we represent the context window $\mathbf{w}_n$ as BOW, the instantiated probability $p_\theta(x_n, \mathbf{w}_n|z_n^{(s)})$ can be therefore derived as:

$$p_\theta(x_n, \mathbf{w}_n|z_n^{(s)}) \propto \exp(\boldsymbol{M}_x z_n^{(s)} + \boldsymbol{b}_x)[x_n] \prod\nolimits_{v=1}^{V} \exp(\boldsymbol{\beta}^{\mathrm{T}}\zeta_n^{(s)} + \boldsymbol{b}_w)[v]^{\mathbf{w}_n[v]} \tag{3.3.8}$$

where $\exp(\boldsymbol{M}_x z_n^{(s)} + \boldsymbol{b}_x)[x_n]$ denotes the x_n-th element of the vector $\exp(\boldsymbol{M}_x z_n^{(s)} + \boldsymbol{b}_x)$.

3.3.4 Loss Function

We are now ready to compute ELBO in Equation 3.3.2 with the specified $q_\phi(z_n|x_n, \mathbf{w}_n)$ and $p_\theta(x_n, \mathbf{w}_n|z_n^{(s)})$ in hand. Our final objective function that needs to be maximized is:

$$\begin{aligned}&\mathcal{L}(\theta, \phi; x_n, \mathbf{w}_n)\\&= \frac{1}{S}\sum_{s=1}^{S} \log p_\theta(x_n, \mathbf{w}_n|\mu_n + \sigma_n \odot \epsilon_n^{(s)}) + \frac{1}{2}\sum_{d=1}^{D}\left(1 + \log\sigma_n[d]^2 - \mu_n[d]^2 - \sigma_n[d]^2\right)\end{aligned} \tag{3.3.9}$$

Here, D denotes the dimension of μ. S denotes the number of sample points required for the computation of the expectation term. The loss function is the negative of the objective function. The learning procedure is summarized in Algorithm 3.1.

Algorithm 3.1: Training of JTW model

Input: pivot words $x_{1:N}$, context windows $\mathbf{w}_{1:N}$, learning rate η, learning rate decay *lrDecay*, maximum iterative number *maxIter*, batch size B, batch number N_B;

Output: learned network parameters θ,ϕ;

Initialize θ, ϕ randomly;
$i \leftarrow 0$, $\eta \leftarrow 0.0005$;
For convenience, define $\mathbf{x}_B = x_{n:n+B}$, $\mathbf{w}_B = \mathbf{w}_{n:n+B}$ as a minibatch;
while θ, ϕ *not converged* and $i < maxIter$ **do**
    Shuffle dataset $x_{1:N}, \mathbf{w}_{1:N}$;
    for 1 **to** N_B **do**
        Generate S samples $\epsilon^{(s)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$;
        Compute gradient $g \leftarrow \nabla_{\theta,\phi}\mathcal{L}(\theta, \phi; \mathbf{x}_B, \mathbf{w}_B)$ according to Equation 3.3.9;
        Update parameters θ, ϕ using gradient g;
    $i \leftarrow i + 1$, $\eta \leftarrow \eta \times lrDecay$;
return θ, ϕ;

3.3.5 Prediction

After training, we are able to map the words to their respective representations using the Encoder part of JTW. The Encoder takes a pivot word together with its context window as an input and outputs the parameters of the variational distribution considered to be the approximated posterior $q_\phi(z|x_n, \mathbf{w}_n)$, which is a Gaussian distribution in our case. The word representations are Gaussian parameters $\{\mu_n, \sigma_n\}$. Because the output of the Encoder is formulated as a Gaussian distribution, the word similarity of two words can be either computed by the KL-divergence between the Gaussian distributions, or by the cosine similarity between their means. We use the Gaussian mean μ to represent a word given its context. The universal representation of a word type can be obtained by averaging the posterior means of all occurrences over the corpus.

3.4 Experimental Setup

Dataset We train the proposed JTW model on the Yelp dataset[2], which is a collection of more than 4 million reviews on over 140k business categories. Although the number of business categories is large, the vast majority of reviews falls into 5 business categories. The top *Restaurant* category consists of more than 40% of

[2] `https://www.yelp.com/dataset/documentation/main`

reviews. The next top 4 categories, *Shopping*, *Beauty&Spas*, *Automotive* and *Clinical* contains about 8%, 6%, 4% and 3% of reviews, respectively. The *Clinical* documents are further filtered by business subcategories defined in Tran and Lee [272], which are recognized as core clinical businesses. This results in 176,733 documents for the *Clinical* category. Because the dataset is extremely imbalanced, simply training the model on the original dataset will likely overfit to the *Restaurant* category. We thus balance the dataset by sampling roughly an equal number of documents from each of the top 5 categories. The vocabulary size is set to 8,000. We use Mallet[3] to filter out stopwords. The final dataset consists of 865,616 documents with a total of 101,468,071 tokens.

Parameter Setting The word semantics are represented as 100-dimensional vectors (i.e., $D = 100$), which is a default configuration for word representations [32, 181]. The number of latent topics is set to 50. It has been previously studied in Kingma and Welling [123] that the number of samples per data point can be set to 1 if the batch size is large, (e.g. > 100). In our experiments, we set the batch size to 2,048 and the number of samples per data point, S, to 1. The context window size is set to 10. Network parameters (i.e., θ, ϕ) are all initialized by a normal distribution with a zero mean and 0.1 variance.

Baselines We compare the JTW model against four baselines:

- **CvMF** [111]. CvMF can be viewed as an extension of GloVe that modifies the objective function by multiplying a mixture of vMFs, whose distance is measured by cosine similarity instead of euclidean distance. The mixture depicts the underlying semantics with which the words could be clustered.
- **Bayesian Skip-Gram (BSG)** [32]. BSG[4] is a probabilistic word-embedding method built on VAE as well, which achieved the state-of-art among other Bayesian word-embedding alternatives [15, 280]. BSG infers the posterior or dynamic embedding given a pivot word and its observed context and is able to learn context-dependent word embeddings.
- **Skip-gram Topical word Embedding (STE)** [248]. STE adapted the commonly known Skip-Gram by associating each word with an input matrix and an output matrix and used the Expectation-Maximization (EM) method

[3] `http://mallet.cs.umass.edu/`
[4] `https://github.com/ixlan/BSG`

with the negative sampling for model parameter inference. For topic generation, they need to evaluate the probability of $p(w_{t+j}|z, w_t)$ for each topic z and each skip-gram $< w_t; w_{t+j} >$, and represent each topic as the ranked list of bi-grams.

- **Mixed Membership Skip-Gram (MMSG)** [69]. MMSG leverages mixed membership modeling in which words are assumed to be clustered into topics and the words in the context of a given pivot word are drawn from the log-bilinear model using the vector representations of the context-dependent topic. Model inference is performed using the Metropolis-Hastings-Walker algorithm with noise-contrastive estimation.

Among the aforementioned baselines, CvMF and BSG only generate word embeddings and do not model topics explicitly. Also, CvMF only maps each word to a single word embedding whereas BSG can output context-dependent word embeddings. Both STE and MMSG can learn topics and topic-dependent embeddings at the same time. However, in STE the topic dependence is stored in the rows of word matrices and the word representations themselves are context-independent. In contrast, MMSG associates each word with a topic distribution; it could produce contextualized word embeddings by summing up topic vectors weighed by the posterior topic distribution given a context. We probe different topic counts and find the best setting for methods with topics or mixtures. In all the baselines, the dimensionality of word embeddings is tuned and finally set to 100.

3.5 Experimental Results

We compare JTW with baselines on both word similarity and word-sense disambiguation tasks for the learned word embeddings. We also present the topic coherence and qualitative evaluation results for the extracted topics. Furthermore, we show that JTW can be easily integrated with deep contextualized word embeddings to improve further the performance of downstream tasks such as sentiment classification.

3.5.1 Word Similarity

The word similarity task [68] has been widely adopted to measure the quality of word embeddings. In the word similarity task, a number of pair-wise words are given. Each pair of words should be assigned with a score that indicates their relatedness. The calculated scores are then compared with the golden scores by means

Table 3.1: Spearman rank correlation coefficient on 7 benchmarks.

Benchmarks	SG	CvMF	BSG	STE	MMSG	JTW (std. dev.)
WS353-SIM	**0.610**	0.597	0.529	0.582	0.579	0.598 (.014)
WS353-ALL	0.571	**0.615**	0.551	0.538	0.558	0.606 (.012)
MEN	0.649	0.632	**0.656**	0.650	0.627	0.653 (.006)
SimLex-999	0.321	0.313	0.271	0.301	0.281	**0.344** (.005)
SCWS	0.620	0.637	**0.652**	0.622	0.624	0.640 (.010)
MTurk771	0.548	0.524	0.555	0.554	**0.596**	0.546 (.010)
MTurk287	0.534	0.517	0.572	**0.641**	0.599	0.639 (.006)
Average	0.550	0.548	0.541	0.555	0.552	**0.575** (.004)

of Spearman rank-order correlation coefficient. Because the word similarity task requires context-free word representations, we aggregate all the occurrences and obtain a universal vector for each word. The distance used for similarity scores is cosine similarity. For STE, we use AvgSimC following Shi et al. [248]. We further make a comparison with the results of the Skip-Gram (SG) model[5], which maps each word token to a single point in an Euclidean space without considering different senses of words. All the approaches are evaluated on the 7 commonly used benchmarking datasets. For JTW, we average the results over 10 runs and also report the standard deviations.

The results are reported in Table 3.1. It can be observed that among the baselines, BSG achieves the lowest score on average, followed by MMSG. Although JTW clearly beats all the other models on SimLex-999 only, it only performs slightly worse than the top model in 5 out of the remaining 6 benchmarks. Overall, JTW gives superior results on average. A noticeable gap can be observed on the Stanford's Contextual Word Similarities (SCWS) dataset where JTW, MMSG and BSG give better results compared with SG, CvMF and STE. This can be explained by the fact that, in SCWS, golden scores are annotated together with the context. However, SG, CvMF and STE can only produce context-independent word vectors. The results show the clear benefit of learning contextualized word vectors. Among the topic-dependent word embeddings, JTW built on VAE appears to be more effective than the PLSA-based STE and the mixed membership model MMSG, achieving the best overall score when averaging the evaluation results across all seven benchmarking datasets. The small standard deviation of JTW indicates that the performance is consistent across multiple runs.

[5] `https://code.google.com/archive/p/word2vec/`

Table 3.2: Accuracy on the lexical substitution task.

Model	CvMF	BSG	STE	MMSG	JTW
Accuracy	0.440	0.453	0.433	0.474	**0.487**

3.5.2 Lexical Substitution

While the word similarity tasks focus more on the general meaning of a word (since word pairs are presented without context), in this section, we turn to the lexical substitution task [269, 307], which was designed to evaluate the word-embedding learning methods regarding their ability to disambiguate word senses. The lexical substitution task can be described by the following scenario: Given a sentence and one of its member words, find the most related replacement from a list of candidate words. As stated in Thater et al. [269], a good lexical substitution should not only capture the relatedness between the candidate word and the original word, but also imply the correctness with respect to the context.

Following Bražinskas et al. [32], we derive the setting from Melamud et al. [176] to ensure a fair comparison between the context-free word embedding methods and the context-dependent ones. In detail, for JTW and BSG, we capture the context of a given word using the BOW representation, and derive the representation of each candidate word taking account of the context. For CvMF and STE, the similarity score is computed using

$$\mathrm{BalAdd}(x, y) = \frac{C\cos(y, x) + \sum_{c=1}^{C}\cos(y, w_c)}{2C}, \tag{3.5.1}$$

where y is the candidate word and x denotes the original word. For MMSG, the original word's representation is calculated as the sum of its associated topic vectors weighed by the word's posterior topical distribution. Given an original word and its context, we choose the candidate word with the highest similarity score. We compare the performance of various models on lexical substitution using the dataset from the SemEval 2007 task 10[6] [173], which consists of 1,688 instances. Because some words have multiple synonyms as annotated in the dataset, we would consider a chosen candidate word as a correct prediction if it hits one of the ground-truth replacements. We report in Table 3.2 the accuracy scores of different methods. Context-sensitive word embeddings generally perform better than context-free alternatives. STE can only learn context-independent word embeddings and hence gives the lowest score. BSG is able to learn context-dependent word embeddings

[6] `http://www.dianamccarthy.co.uk/task10index.html`

and outperforms CvMF. Among the joint topic and word embedding learning methods, STE performs the worst, showing that associating each word with two matrices and learning topic-dependent word embeddings based on PLSA appear to be less effective. Both JTW and MMSG show superior performances compared to BSG. JTW outperforms MMSG because JTW also models the generation of pivot word in addition to context words and the VAE framework for parameter inference is more effective than the annealed negative contrastive estimation used in MMSG.

3.5.3 Topic Coherence

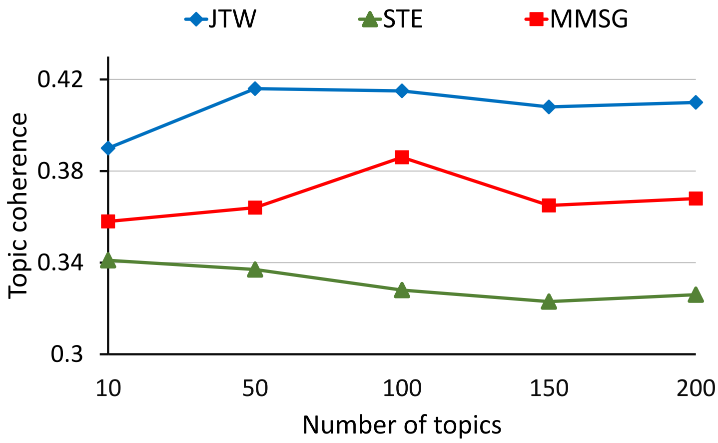


Figure 3.2: Topic coherence scores versus the number of topics.

Because only STE and MMSG can jointly learn topics and word embeddings among the baselines, we compare our proposed JTW with these two models in terms of topic quality. The evaluation metric we employed is the topic coherence metric proposed in Röder et al. [231]. The metric extracts co-occurrence counts of the topic words in Wikipedia using a sliding window of size 110. For each top word a vector is calculated whose elements are the normalized point-wise mutual information between the word and every other top word. Given a topic, the arithmetic mean of all vector pairs' cosine similarity is treated as the coherence measure. We calculate the topic coherence score of each extracted topic based on its associated top ten words using Palmetto[7] [233]. The topic coherence results with the topic number varying between 10 and 200 are plotted in Figure 3.2. The graph shows that JTW scores the highest under all the topic settings. It gives the best coherence score of 0.416 at 50 topics, and gradually flattens with the increasing number of topics. MMSG exhibits an upward trend up to 100 topics, and drops to 0.365 when the topic number is

[7] `https://github.com/dice-group/Palmetto`

Table 3.3: Example topics discovered by JTW and MMSG, each topic is represented by the top 10 words sorted by their likelihoods. The topic labels are assigned manually. Semantically less coherent words are highlighted by *italics*.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
Food	*Shopping*	*Beauty*	*Automotive*	*Clinical*
		JTW		
good	great	hair	car	compassionate
food	friendly	recommend	*told*	caring
chicken	service	highly	phone	personable
place	staff	place	called	courteous
pizza	shop	experience	care	therapy
love	clean	fabulous	vehicle	competent
cheese	helpful	great	*time*	knowledgeable
salad	nice	nail	BMW	passionate
red	amazing	nails	insurance	physician
delicious	customer	awesome	wanted	respectful
		MMSG		
food	friendly	massage	place	therapy
service	staff	spa	service	physical
great	great	back	*time*	pain
good	helpful	great	*back*	*back*
place	service	*time*	customer	*massage*
friendly	clean	good	car	recommend
staff	place	massages	*people*	great
nice	nice	facial	good	therapist
back	store	*therapist*	money	*work*
prices	super	body	*give*	highly

set to 150. STE undergoes a gradual decrease and then stabilizes with the topic number beyond 150.

3.5.4 Extracted Topics

We present in Table 3.3 the example topics extracted by JTW and MMSG. It can be easily inferred from the top words generated by JTW that Topic 1 is related to '*Food*', whereas Topic 5 is about the '*Clinical Service*', which is identified by the words '*caring*' and '*physician*'. It can also be deduced from the top words that Topic 2, 3 and 4 represent '*Shopping*', '*Beauty*' and '*Automotive*', respectively. In contrast, topics produced by MMSG contain more semantically less coherent words as highlighted by italics. For example, Topic 1 in MMSG contains words relating to both food and staff. This might be caused by the fact that, in MMSG, training is performed as a two-stage process by first assigning topics to words using Gibbs Sampling then estimating the topic vectors and word vectors from word co-occurrences and topic assignments via maximum likelihood estimator. This is equivalent to a topic model with parameterized word embeddings. Conversely, in JTW, latent variables in the generative process are recognized as word representations. Parameters reside in the generative network, and are inferred by the VAE. No extra parameters are introduced to encode the words. Therefore, the topics extracted tend to be more identifiable.

3.5.5 Visualization of Word Semantics

The extracted topics allow the visualization of word semantics, which facilitates the interpretation [142, 303]. In JTW, a word's semantic meanings can be interpreted as a distribution over the discovered latent topics. This is achieved by aggregating all the contextualized topical distribution of a particular word throughout the corpus. Meanwhile, when a word is placed under a specific context, its topical distribution can be directly transformed from its contextualized representation. We chose three words—'*plastic*', '*bar*' and '*patient*'—to illustrate the polysemous nature of them. To further demonstrate their context-dependent meanings, we also visualize the topic distribution of the following three sentences: (1) *Effective patient care requires clinical knowledge and understanding of physical therapy*; (2) *Restaurant servers require patient temperament*; (3) *You have to bring your own bags or boxes but you can also purchase plastic bags.* The topical distribution for the pivot words and the three example sentences are shown in Figure 3.3.

We can deduce from the overall distributions that the semantic meaning of '*plastic*' distributes almost equally on two topics, '*shopping*' and '*beauty*', while the meaning of '*bar*' is more prominent on the '*food*' and '*shopping*' topics. '*Patient*' has a strong connection with the '*clinical*' topic, though it is also associated with the '*food*' topic. When considering a specific context about the patient care, Sentence 1 has its topic distribution peaked at the '*clinical*' topic. Sentence 2 also contains the word '*patient*', but it now has its topic distribution peaked at '*food*'. Sentence 3 mentioned '*plastic bags*' and its most prominent topic is '*shopping*'. These results show that JTW can indeed jointly learn latent topics and topic-specific word embeddings.

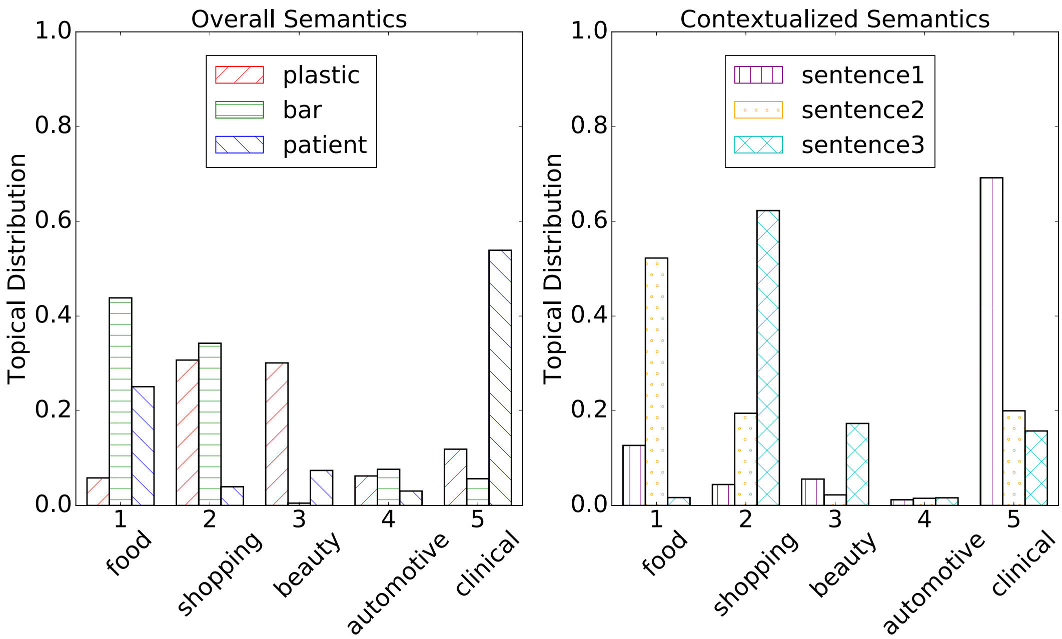


Figure 3.3: The overall topical distributions and contextualized topical distributions of the example words and the contextualized topical distribution of three example sentences. Note that the x-axis denotes the five example topics shown in Table 4.

3.5.6 Integration with Deep Contextualized Word Embeddings

Advances in deep contextualized word representation learning have significantly impacted natural language processing [141]. Different from traditional word embedding learning methods such as Word2Vec or GloVe, where each word is mapped to a single vector representation, deep contextualized word representation learning methods are typically trained by language modelling and generate a different word vector for each word depending on the context in which it is used. A notable work is ELMo [213], which is commonly regarded as the pioneer for deriving deep contextualized word embeddings [60]. ELMo calculates the weighted sum of different

Table 3.4: Results on the 5-class sentiment classification by 10-fold cross validation on the Yelp reviews.

Model	Criteria			
	Precision	Recall	Macro-F1	Micro-F1
JTW	0.5713±.021	0.5639±.014	0.5599±.016	0.7339±.015
ELMo	0.6091±.005	0.6053±.001	0.6056±.002	0.7610±.005
BERT	0.6293±.014	0.5952±.006	0.6041±.012	0.7626±.005
JTW-ELMo	0.6286±.008	**0.6110±.004**	**0.6168±.008**	0.7783±.004
JTW-BERT	**0.6354±.014**	0.6081±.009	0.6045±.014	**0.7806±.005**

layers of a multi-layered BiLSTM-based language model, using the normalized vector to represent the corresponding word. Another more recent work is BERT [60]. In contrast to ELMo, BERT [60] was proposed to apply the bidirectional training of Transformer to masked language modelling. Because of its capability of effectively encoding contextualized knowledge from massive external corpora in word embeddings, BERT has refreshed the state-of-art results on a number of NLP tasks.

While Word2Vec/GloVe and ELMo/BERT represent the two opposite extremes in word embedding learning, with the former learning a single vector representation for each word and the latter learning a separate vector representation for each occurrence of a word, our proposed JTW sits in the middle that it learns different word vectors depending on which topic a word is associated with. Nevertheless, we can incorporate ELMo/BERT embeddings into JTW. This is achieved by replacing the BOW input with the pre-trained ELMo/BERT word embeddings in the Encoder-Decoder architecture of JTW, making the resulting word embeddings better at capturing semantic topics in a specific domain. More precisely, the training objective is switched to the cosine value of half the angle between the input ELMo/BERT vector and decoded output vector formulated as:

$$p_\theta(x_n, \mathbf{w}_n|z_n^{(s)}) \propto \cos(\frac{1}{2}\arccos(\frac{x_n^\top \cdot x_n^{(p)}}{\|x_n\|\|x_n^{(p)}\|})) \prod\nolimits_{c=1}^{C} \cos(\frac{1}{2}\arccos(\frac{w_{n,c}^\top \cdot w_{n,c}^{(p)}}{\|w_{n,c}\|\|w_{n,c}^{(p)}\|})), \tag{3.5.2}$$

where $x_n^{(p)}$ and $w_{n,c}^{(p)}$ are the reconstructed representations generated from $z_n^{(s)}$ by Equation 3.3.5 and Equation 3.3.7, respectively. Recall that, the input to the model has been encoded by pre-trained word vectors (e.g., 300-dimensional vectors). Our training objective is to make the reconstructed $x_n^{(p)}$ and $w_{n,c}^{(p)}$ as close as possible to their original input word embeddings. The difference is measured by the angle between the input and the output vectors. Normalized ELMo/BERT vectors can be transformed to the polar coordinate system with trigonometric functions, which forms a probability distribution by

$$\int_0^\pi \frac{1}{2}\cos\frac{\theta}{2}\mathrm{d}\theta = 1, \tag{3.5.3}$$

and the function is monotone to the similarity between the input ELMo/BERT embeddings and the reconstructed output embeddings, which reaches its peak when $x_n = x_n^{(p)}$ (i.e., $\theta = 0$). Therefore, we are able to replace Equation 3.3.8 with Equation 3.5.2 when an ELMo/BERT is attached. The input vectors of the Encoder are then the embeddings produced by ELMo/BERT, and the Decoder output is the reconstructed word embeddings aligned with the input.

We resort to the sentiment classification task on Yelp and compare the performance of JTW, ELMo and BERT[8], and the integration of both, JTW-ELMo and JTW-BERT, by 10-fold cross validation. In all the experiments, we fine-tune the models on the training set consisting of 90% documents sampled from the dataset described in Section 3.4 and evaluate on the 10% data that serves as the test set. We employ the further pre-training scheme [261] that different learning rates are applied to each layer, and slanted triangular learning rates are imposed across epochs when adapting the language model to the training corpus [99]. The classifier used for all the methods is an attention hop over a BiLSTM with a softmax layer. The ground truth labels are the five-scale review ratings included in the original dataset. The 5-class sentiment classification results in precision, recall, macro-F1 and micro-F1 scores are reported in Table 3.4.

It can be observed from Table 3.4 that a sentiment classifier trained on JTW-produced word embeddings gives worse results compared with that using the deep contextualized word embeddings generated by ELMo or BERT. Nevertheless, when integrating the ELMo or BERT front-end with JTW, the combined model, JTW-ELMo and JTW-BERT, outperforms the original deep contextualized word representation models, respectively. It has been verified by the paired t-test that JTW-ELMo outperforms ELMo and BERT at the 95% significance level on Micro-F1. The results show that the proposed JTW is flexible and can be easily integrated with pre-trained contextualized word embeddings to capture the domain-specific semantics better compared to directly fine-tuning the pre-trained ELMo or BERT on the target domain, hence leading to improved sentiment classification performance.

3.6 Summary

Driven by the motivation that combining word embedding learning and topic modelling can mutually benefit each other, we propose a probabilistic generative framework that can jointly discover more semantically coherent latent topics from the global context and learn topic-specific word embeddings, which naturally addresses the problem of word polysemy. Experimental results verify the effectiveness of the model on word similarity evaluation and word sense disambiguation. Furthermore, the model can discover latent topics shared across documents, and the encoder of JTW can generate the topical distribution for each word. This enables an intuitive understanding of word semantics. We have also shown that our proposed JTW can be easily integrated with deep contextualized word embeddings to improve the

[8] `https://github.com/google-research/bert`

performance of downstream tasks further.

Chapter 4

A Neural Opinion Dynamics Model for Temporal Stance Prediction

Chapter Abstract

In this chapter, we model users' posting behaviour on social media as a temporal point process to jointly predict the posting time and the stance label of the next tweet, given a user's historical tweet sequence and tweets posted by their neighbours. Opinion prediction on Twitter is challenging since users' opinions are not only volatile but also changeable over time due to the influences from their neighbours on social networks or arguments they encounter that undermine their beliefs. To tackle this, we design a topic-driven attention mechanism to capture the dynamic topic shifts in the neighbourhood context. In what follows, we first introduce the background of neural opinion dynamics. Then we proceed to the network structure. Finally, we report experimental results on posting time prediction and stance prediction. The proposed model showed higher accuracy compared to several competitive baselines.

4.1 Introduction

Social media platforms allow users to express their opinions online towards various subject matters. Despite much progress in sentiment analysis in social media, the prediction of opinions, however, remains challenging. Opinion formation is a complex process. An individual's opinion could be influenced by their own prior belief, their social circles and external factors. Existing studies often assume that socially connected users hold similar opinions. Social network information is integrated with user representations via weighted links and encoded using neural networks with attention or Graphical Convolutional Networks (GCNs) [40, 138]. This strand of work, including [41, 59, 331], leverages both the chronological tweet sequence and social networks to predict users' opinions.

The majority of previous work requires a manual segmentation of a tweet sequence into equally-spaced intervals based on either tweet counts or time duration. Models trained on the current interval are used to predict users' opinions in the next interval. However, we argue that such a manual segmentation may not be appropriate since users post tweets at different frequencies. Also, the time interval between two consecutively published tweets by a user is important to study the underlying opinion dynamics system and hence should be treated as a random variable.

Inspired by the multivariate Hawkes process [1, 64], we propose to model a user's posting behaviour by a temporal point process that when user u posts a tweet d at time t, they need to decide on whether they want to post a new topic/opinion, or post a topic/opinion influenced by past tweets either posted by other users or by themselves. We thus propose a neural temporal opinion model to jointly predict the time when the new post will be published and its associated stance. Instead of using the fixed formulation of the multivariate Hawkes process, the intensity function of the point process is automatically learned by a gated recurrent neural network. In addition, one's neighbourhood context and the topics of their previously published tweets are also taken into account for the prediction of both the posting time and stance of the next tweet.

4.2 Related Work

The prediction of real-time stances on social media is challenging, partly caused by the diversity and fickleness of users [5]. A line of work mitigated the problem by taking into account the homophily that users are similar to their friends [86, 174]. For example, Chen et al. [40] gauged a user's opinion as an aggregated stance of their neighbourhood users. Linmei et al. [151] took a step further by exploiting the extracted topics, which discern a user's focus on neighbourhood tweets. Related works in this strand also include the application of GCNs, with which the social relationships are leveraged to enrich the user representations [59, 138].

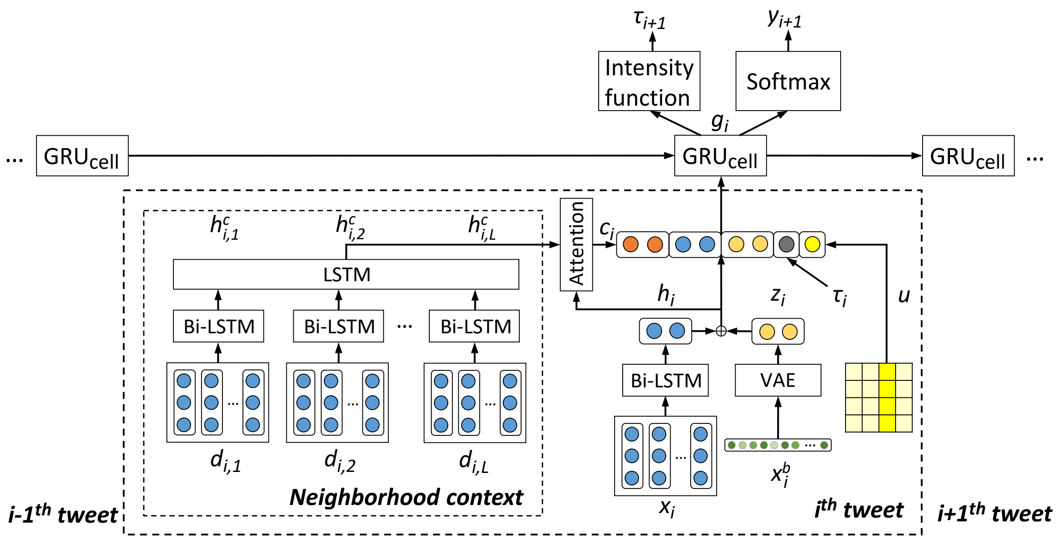


Figure 4.1: Overview of the Neural Temporal Opinion Model.

On the other hand, several works have utilized the chronological order of tweets. Chen et al. [41] presented an opinion tracker that predicts a stance every time a user publishes a tweet, whereas [331] extended the previous work by introducing a topic-dependent attention. Shrestha et al. [249] considered diverse social behaviors and jointly forecast them through a hierarchical neural network. Zhao et al. [327] employed Poisson factorization to deal with trunks of streaming documents. However, the aforementioned work requires a manual segmentation of a tweet sequence. Furthermore, they are unable to predict when a user will next publish a tweet and what its associated stance is. These problems can be addressed using the Hawkes process [89], which has been successfully applied to event tracking [257], rumor detection [4, 163, 336] and retweet prediction [129]. A combination of the Hawkes process with recurrent neural networks, called Recurrent Marked Temporal Pointed Process (RMTPP), was proposed to automatically capture the influence of the past events on future events, which shows promising results on geolocation prediction [64]. Benefiting from the flexibility and scalability of neural networks, several work has been done in this vein including event sequence prediction [175] and failure prediction [298]. Our work is partly inspired by RMTPP, but departs from the previous work by jointly considering users' social relations and topical attentions for stance prediction on social media.

4.3 Neural Temporal Opinion Model

We present the overall architecture in Figure 4.1. The input to the model at time step i consists of user's own tweet x_i, bag-of-word representation x_i^b, time interval τ_i between the $i-1^{th}$ tweet and the i^{th} tweet, user embedding u, and neighbours' tweet queue $\{d_{i,1}, d_{i,2}, \ldots, d_{i,L}\}$. At first, a Bi-LSTM layer is applied to extract features from input tweets. Then the neighbourhood tweets are processed by a stacked Bi-LSTM/LSTM layer for the extraction of neighbourhood context, which is fed into an attention module queried by the user's own tweet h_i and topic z_i. The output of the attention module is concatenated with tweet representation, time interval τ_i, user representation u, and topic representation z_i, which is encoded from x_i^b via a Variational Autoencoder (VAE). Finally, the combined representation is sent to a GRU cell, whose hidden state participates in computing the intensity function and the softmax function, for the prediction of the posting time interval and the stance label of the next tweet. In the following, we elaborate the model in more details:

Tweet representation: Words in tweets are mapped to pre-trained word embeddings [17][1], which is specially trained for tweets. Then Bi-LSTM is used to generate the tweet representation.

Topic extraction: The topic representation z_i in Figure 4.1 captures the topic focus of the i^{th} tweet. It is learned by VAE [123], which approximates the intractable true posterior by optimising the reconstruction error between the generated tweet and the original tweet. Specifically, we convert each tweet to the bag-of-word format weighted by term frequency, x_i^b, and feed it to two inference neural networks defined as f_{μ_ϕ} and f_{Σ_ϕ}. These generate the mean and variance of a Gaussian distribution from which the latent topic vector z_i is sampled. Then the approximated posterior would be $q_\phi(z_i|x_i^b) = \mathcal{N}(z_i|f_{\mu_\phi}(x_i^b), f_{\Sigma_\phi}(x_i^b))$. To generate the observation $\tilde{x}_i^b$ conditional on the latent topic vector z_i, we define the generative network as $p_\varphi(x_i^b|z_i) = \mathcal{N}(x_i^b|f_{\mu_\varphi}(z_i)), f_{\Sigma_\varphi}(z_i))$. The reconstruction loss for the tweet x_i^b is then:

$$\mathcal{L}_x = \mathbb{E}_{q_\phi(z_i|x_i^b)}[\log p_\varphi(x_i^b|z_i)] - \mathrm{KL}(q_\phi(z_i|x_i^b)||p(z_i)) \tag{4.3.1}$$

Neighbourhood Context Attention: To capture the influence from the neighbourhood context, we first input the neighbours' recent L tweets to an LSTM in a temporal ascending order. The output of the LSTM is weighed by the attention

[1] `https://github.com/cbaziotis/datastories-semeval2017-task4`

signals queried by the user's i^{th} tweet and topic:

$$c_i = \sum_{l=1}^{L} \alpha_l h^c_{i,l} \tag{4.3.2}$$

$$\alpha_l \propto \exp([h_i^\intercal, z_i^\intercal]\tanh(W_h h^c_{i,l} + W_z z^c_{i,l})) \tag{4.3.3}$$

where $\{h^c_{i,1}, h^c_{i,2}, \dots, h^c_{i,L}\}$ denotes the hidden state output of each tweet $d_{i,l}$ in the neighbourhood context, $z^c_{i,l}$ denotes the associated topic, h_i is the representation of the user's own tweet at time step i, and both W_h and W_z are weight matrices.

We use this attention mechanism to align the user's tweet to the most relevant part in the neighbourhood context. Our rationale is that a user would attend to their neighbours' tweets that discuss similar topics. The attention output c_i is then concatenated with a user's own tweet h_i and the extracted topic z_i. We further enrich the representation with the elapsed time τ_i between the posting time of the current tweet and the last posted tweet, and add a randomly initialised user vector u to distinguish the user from others. The final representation is passed to a GRU cell for the joint prediction of the posting time and stance label of the next tweet.

Temporal Point Process: The goal of NTOM is to forecast the time gap till the next post, together with the stance label. Instead of modelling the time interval value based on regression analysis, we use the GRU [46] to simulate the temporal point process.

At each time step, the combined representation $[c_i, h_i, z_i, \tau_i, u]$ is input to the GRU cell to iteratively update the hidden state taking into account the influence of previous tweets:

$$g_i = f_{GRU}(g_{i-1}, c_i, h_i, z_i, \tau_i, u) \tag{4.3.4}$$

where g_i is the hidden state of GRU cell. Given g_i, the intensity function is formulated as:

$$\lambda^*(t) = \lambda(t|\mathcal{H}_i) = \exp(b_\lambda + v_\lambda^\intercal g_i + w_\lambda t) \tag{4.3.5}$$

Here, $\mathcal{H}_i$ summarises all the tweet histories up to tweet i, b_λ denotes the base density level, the term $v_\lambda^\intercal g_i$ captures the influence from all previous tweets and $w_\lambda t$ denotes the influence from the instant interval. The likelihood that the next tweet will be posted at the next interval τ given the history is:

$$f^*(\tau) = \lambda^*(\tau) \exp\big(-\int_0^\tau \lambda^*(t)dt\big) \tag{4.3.6}$$

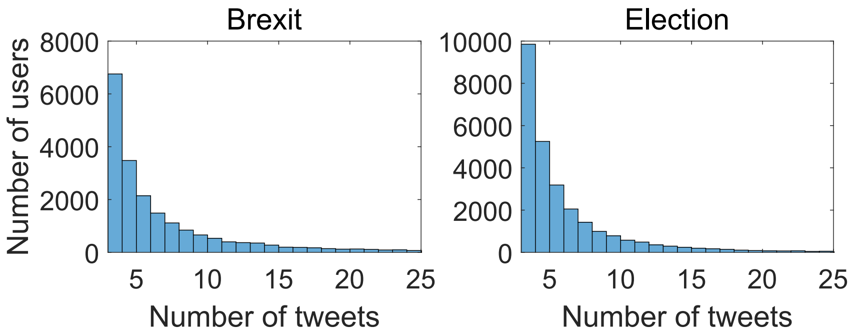


Figure 4.2: Number of users versus number of tweets.

The expectation for the occurrence of the next tweet can be estimated using:

$$\hat{\tau}_{i+1} = \int_0^\infty \tau \cdot f^*(\tau) d\tau \tag{4.3.7}$$

Loss: We expect the predicted interval to be close to the actual interval as much as possible by minimising the Gaussian penalty function:

$$\mathcal{L}_{time} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(\tau_{i+1} - \hat{\tau}_{i+1})^2}{2\sigma^2}\right) \tag{4.3.8}$$

For the stance prediction we employ the cross-entropy loss denoted as $\mathcal{L}_{stan}$. The final objective function is computed as:

$$\mathcal{L} = \eta\mathcal{L}_x + \beta\mathcal{L}_{time} + \gamma\mathcal{L}_{stan} \tag{4.3.9}$$

where η, β and γ are coefficients determining the contribution of various loss functions.

4.4 Experimental Setup

We perform experiments on two publicly available Twitter datasets[2] [331] on Brexit and US election. The Brexit dataset consists of 363k tweets with 31.6%/29.3%/39.1% supporting/opposing/neutral tweets towards Brexit. The Election dataset consists of 452k tweets with 74.2%/20.4%/5.4% supporting/opposing/neutral tweets towards Trump. We filter out users who posted less than 3 tweets and are left with $20,914$ users in Brexit and $26,965$ users in Election. We plot in Figure 4.2 the number of users versus the number of tweets and found that over 81.6% users have published fewer than 7 tweets, we, therefore, set the maximum length of the tweet sequence of

[2] `https://github.com/somethingx01/TopicalAttentionBrexit`

each user to 7. For users who have published more than 7 tweets, we split their tweet sequence into multiple training sequences of length 7 with an overlapping window size of 1. For each user, we use 90% of their tweets for training and 10% (round up) for testing.

The settings are $\eta = 0.2$, $\beta = 0.4$ and $\gamma = 0.4$. We set the topic number to 50 and the vocabulary size to 3k for the tweet bag-of-words input to VAE. The mini-batch size is 16. We use Adam optimizer with a learning rate of 0.0005 and a learning rate decay of 0.9. The evaluation metrics are accuracy for stance prediction and Mean Squared Error (MSE) for posting time prediction. The results are compared against the following baselines:

- CSIM_W [41] gauges the social influence by an attention mechanism for the prediction of the user sentiment of the next tweet.
- NOD [331] takes into account the neighborhood context and pre-extracted topics for tweet stance prediction.
- LING+GAT [59] places a GCN variant over linguistic features to extract node representations. Tweets are aggregated by users for user-level prediction.

We also perform an ablation study on our model by removing the topic extraction component ($\text{NTOM}_{\text{-VAE}}$) or removing the neighbourhood context component ($\text{NTOM}_{\text{-context}}$). In addition, to validate that NTOM does benefit from point process modelling and can better forecast the time and stance of the next tweet, we remove the intensity function (i.e. no Eq. (5)-(7)) and directly use vanilla RNN and its variants including LSTM and GRU to predict the true time interval. Furthermore, to investigate if is is more beneficial to use GCN to encode the neighbourhood context, we learn tweet representation using GCN[3] [87], which preserves high-order influence in social networks through convolution. As in [138], we use a 2-hop GCN and denote the variant as $\text{NTOM}_{\text{-GCN}}$. For the Brexit dataset, MSE is measured in hours, while for the Election dataset it is measured in minutes due to the intensive tweets published within two days.

4.5 Results

We report in Table 4.1 the stance prediction accuracy and MSE scores of predicted posting time. Compared to baselines, NTOM consistently achieves better performance on both datasets, showing the benefit of modelling the tweet posting sequence as a temporal point process. In the second set of experiments, we study the effect of temporal process modelling. The results verify the benefit of using the intensity

[3] `https://github.com/williamleif/GraphSAGE`

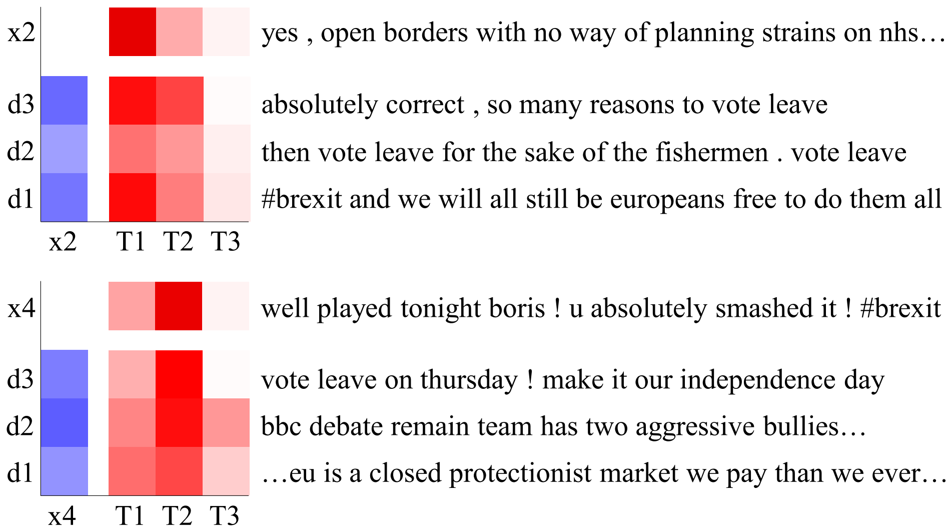
Model	Brexit		Election	
	Acc.	MSE	Acc.	MSE
CSIM_W	0.653	–	0.656	–
NOD	0.675	–	0.690	–
LING+GAT	0.692	–	0.704	–
RNN	0.636	7.81	0.659	9.62
LSTM	0.677	3.37	0.683	4.51
GRU	0.691	2.80	0.693	3.92
NTOM_VAE	0.697	2.67	0.705	4.01
NTOM_context	0.665	3.34	0.682	4.78
NTOM_GCN	0.680	2.65	0.706	4.29
NTOM	**0.713**	**2.59**	**0.715**	**3.70**

Table 4.1: Stance prediction accuracy and Mean Squared Errors of predicted posting time on the Brexit and Election datasets.

function, with at least a 2% increase in accuracy and 0.2 decrease in MSE compared with vanilla RNN and its variants. In the ablation study, removing the neighbourhood context component caused the largest performance decline compared to other components, verifying the importance of social influence in opinion prediction. Removing either VAE (for topic extraction) or intensity function (using only GRU) results in slight drops in stance prediction and more noticeable performance gaps in time prediction. It can also be observed that using GCN to model higher-order influence in social networks does not bring any benefits, possibly due to extra noise introduced to the model.

To investigate the effectiveness of the context attention that is queried by topics, we first select some example topics from the topic-word matrix in VAE. The label of each topic is manually assigned based on its associated top 10 words. Then we display a tweet's topic distribution together with its neighbourhood tweets' topic distribution. We also visualize the attention weights assigned to the 3 neighbourhood tweets.

Figure 4.3 illustrates the example topics, topic distribution and attention signals towards context tweets. Here, x_2 and x_4 denote a user's 2^{nd} and 4^{th} tweets respectively. The most recent 3 neighbourhood tweets are denoted as d_1, d_2, d_3. Blue in the leftmost separate column denotes the attention weights, and each row on top of $T1, T2$ and $T3$ denotes the topic distribution. It can be observed that the user's concerned topic shifts from *immigration* to *Boris Johnson* in 2 time steps. The drift also appears in the neighbour's tweets. Higher attention weights are assigned to the neighbour's tweets which share similar topical distribution as the user. We can thus



	Topic	**Top words**
T1	immigration	immigration, stop, free, work, change, countries, immigrants, migrants, migration, open
T2	Boris Johnson	Boris, live, Johnson, politics, sturgeon, TV, Nicola, morning, takebackcontrol, guy
T3	vote remain	voteremain, strongerin, Cameron, eureferendum, David, inorout, pm, eudebate, osborne, positive

Figure 4.3: Distribution over 3 topics and attention signals on 3 neighbourhood tweets, respectively in 2-time steps. Topics are labelled based on the top 10 words.

infer that the topic vector does help select the most relevant neighbourhood tweet.

4.6 Summary

We have proposed a novel Neural Temporal Opinion Model (NTOM) to address users' changing interests and dynamic social context. We model users' tweet posting behaviour based on a temporal point process for the joint prediction of the posting time and stance label of the next tweet. Experimental results verify the effectiveness of the model. Furthermore, the visualisation of the topics and attention signals shows that NTOM captures the dynamics in the focused topics and contextual attention.

Chapter 5

Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection

Chapter Abstract

In this chapter, we cope with challenges in dialogue emotion detection as it often requires the identification of thematic topics, the relevant commonsense knowledge, and the intricate transition patterns between the affective states to capture the holistic pattern underlying a conversation. We first design a topic-augmented language model (LM) with an additional layer specialized for topic detection. The topic-augmented LM is then combined with commonsense statements derived from a knowledge base based on the dialogue contextual information. Finally, a transformer-based encoder-decoder architecture fuses the topical and commonsense information and performs the emotion label sequence prediction. The model has been experimented on four datasets in dialogue emotion detection, demonstrating its superiority empirically over the existing state-of-the-art approaches. Quantitative and qualitative results show that the model can discover topics which help in distinguishing emotion categories.

5.1 Introduction

The abundance of dialogues extracted from online conversations and TV series provides an unprecedented opportunity to train models for automatic emotion detection, which are essential for developing empathetic conversational agents or chatbots for psychotherapy [37, 100, 113, 314]. However, capturing the contextual semantics of personal experience described in one's utterance is challenging. An instance is "*I just passed the exam*" where the emotion can be either *happy* or *sad* depending on the expectation of the subject. There are strands of works utilizing the dialogue context to enhance the utterance representation [113, 169, 314], where recurrent units handled influences from historical utterances, and attention signals were further introduced to intensify the positional order of the utterances.

However, despite the salient progress made by the aforementioned methods, detecting emotions in dialogues is still challenging due to how emotions are expressed and how their meaning can vary based on the topic discussed, as well as the implicit knowledge shared between participants. Figure 5.1 gives an example of how topics and background knowledge could impact the mood of interlocutors. Normally, dialogues around specific topics carry specific language patterns [241], affecting not only the utterance's meaning but also the particular emotions conveyed by specific expressions. Dialogue emotion detection methods so far did not put emphasis on modelling these holistic properties of dialogues (i.e., conversational topics and tones). Consequently, they were fundamentally limited in capturing the affective states of interlocutors related to the particular themes discussed. Besides, emotion and topic detection heavily relies on leveraging underlying commonsense knowledge shared between interlocutors. Although there have been attempts in incorporating it, such as the Ghosal et al. [76]'s COSMIC, existing approaches do not perform fine-grained extraction of relevant information based on both the topics and the emotions involved.

Recently, the Transformer architecture [277] has empowered language models to transfer large quantities of data to low-resource domains, making it viable to discover topics in conversational texts. On top of this, we propose to add an extra layer to the pre-trained language model to model the latent topics, which are learned by fine-tuning dialogue datasets to alleviate the data sparsity problem. Inspired by the success of Transformers, we use the Transformer Encoder-Decoder structure to perform the Seq2Seq prediction in which an emotion label sequence is predicted given an utterance sequence (i.e., each utterance is assigned with an emotion label). We posit that the dialogue emotion of the current utterance depends on the historical

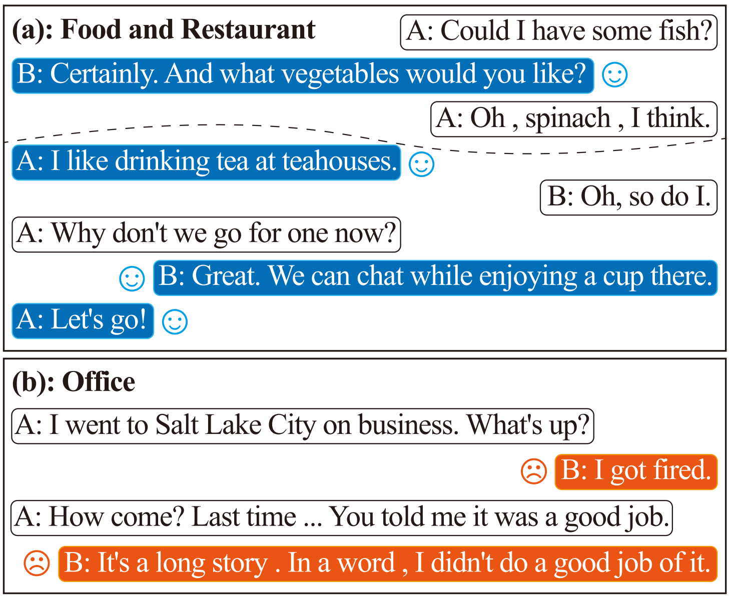


Figure 5.1: Utterances around particular topics carry specific emotions in the DailyDialog dataset. Utterances carrying *positive* (smiling face) or *negative* (crying face) emotions are highlighted in colour. Other utterances are labeled as '*Neutral*'.

dialogue context and the predicted emotion label sequence for the past utterances. We leverage attention mechanisms and gating mechanisms to incorporate various commonsense knowledge retrieved by multiple approaches[1].

5.2 Related Work

Dialogue Emotion Detection Majumder et al. [169] recognized the importance of dialogue context in dialogue emotion detection. They used a Gated Recurrent Unit (GRU) to capture the global context which is updated by the speaker ad-hoc GRUs. At the same time, Jiao et al. [113] presented a hierarchical neural network model that comprises two GRUs for the modelling of tokens and utterances respectively. Zhang et al. [314] explicitly modelled the emotional dependencies on context and speakers using a Graph Convolutional Network (GCN). Meanwhile, Ghosal et al. [75] extended the prior work [169] by taking into account the intra-speaker dependency and relative position of the target and context within dialogues. Memory networks have been explored in [114] to allow bidirectional influence between utterances. A similar idea has been explored by Li et al. [144]. While the majority of works have been focusing on textual conversations, Zhong et al. [328] enriched ut-

[1]Code and trained models are available at `http://github.com/something678/TodKat`

terances with concept representations extracted from the ConceptNet [256]. Ghosal et al. [76] developed COSMIC which exploited Atomic [237] for the acquisition of commonsense knowledge. Unlike the aforementioned approaches, we propose a topic-driven and knowledge-aware model built on a Transformer Encoder-Decoder structure for dialogue emotion detection. The proposed Seq2Seq structure is identical to KET [328] in that emotions are predicted taking into account both the historical utterances and emotions and that a decoder is employed to handle these contexts.

Latent Variable Models for Dialogue Context Modelling Latent variable models, normally described in their neural variational inference form named Variational Autoencoder (VAE) [123], have been studied extensively on learning thematic representations of individual documents [179, 226, 258]. They have been successfully applied to dialogue generation for the benefit of capturing thematic characteristics while retaining a level of flexibility between conversations. This line of work, including those based on hierarchical recurrent VAEs [205, 241, 310] and conditional VAEs [71, 247, 254], encode each utterance with historical latent codes and autoregressively reconstruct the input sequence. On the other side, pre-trained language models are used as embedding inputs to VAE-based models [7, 207]. More recent work by Li et al. [140] employs BERT and GPT-2 as the encoder-decoder structure of VAE. However, these models have to be either trained from scratch or built upon pre-trained embeddings. They are therefore not fitting the low-resource setting of dialogue emotion detection, and cannot benefit from the co-occurrence pattern of utterances within dialogues.

Knowledge Base and Knowledge Retrieval ConceptNet [256] captures commonsense concepts and relations as a semantic network, which encompasses the spatial, physical, social, temporal, and psychological aspects of everyday life. In more recent work, Sap et al. [237] built Atomic, a knowledge graph centred on events rather than entities. Owing to the expressiveness of events and ameliorated relation types, using Atomic achieved competitive results against human evaluation in the task of If-Then reasoning.

Alongside the development of knowledge bases, recent years have witnessed the thriving of new methods for training language models from large-scale text corpora as an implicit knowledge base. As shown in [214], pre-trained language models perform well in recalling relational knowledge involving triplet relations about entities. Bosselut et al. [28] proposed COMmonsEnse Transformers (Comet) which learns to generate commonsense descriptions in natural language by fine-tuning pre-trained language models on existing commonsense knowledge bases such as ATOMIC. Compared with extractive methods, language models fine-tuned on knowledge bases have a distinct advantage of being able to generate knowledge for unseen events, which is of great importance for tasks which require the incorporation of commonsense knowledge such as emotion detection in dialogues.

5.3 Methodology

5.3.1 Problem Setup

A dialogue is defined as a sequence of utterances $\{x_1, x_2, \ldots, x_N\}$, which is annotated with a sequence of emotion labels $\{y_1, y_2, \ldots, y_N\}$. Our goal is to develop a model that can assign the correct label to each utterance. As for each utterance, the raw input is a token sequence, i.e., $x_n = \{w_{n,1}, w_{n,2}, \ldots, w_{n,M_n}\}$ where M_n denotes the length of an utterance. We address this problem using the Seq2Seq framework [262]. In the Seq2Seq framework, the model consecutively consumes an utterance x_n and predicts the emotion label y_n based on the earlier utterances seen so far and their associated predicted emotion labels. The joint probability of emotion labels for a dialogue is:

$$P_\theta(y_{1:N}|x_{1:N}) = \prod_{n=1}^{N} P_\theta(y_n|x_{\leq n}, y_{<n}) \tag{5.3.1}$$

It is worth mentioning that the subsequent utterances are unseen to the model at each predictive step. Learning is performed via optimising the log-likelihoods of predicted emotion labels.

The proposed topic-driven and knowledge-aware transformer consists of two main components, the topic-driven language model fine-tuned on dialogues and the knowledge-aware transformer for emotion label sequence prediction for a given dialogue. In what follows, we will describe each of the components in turn.

5.3.2 Topic Representation Learning

We propose to insert a topic layer into an existing language model and fine-tune the pre-trained language model on the conversational text for topic representation learning. Topic models, often formulated as latent variable models, play a vital role in dialogue modelling [241] due to the explicit modelling of 'high-level syntactic features such as style and topic' [31]. Despite the tremendous success of applying topic modelling in dialogue generation [71, 247, 254], there is scarce work exploiting latent variable models for dialogue emotion detection. To this end, we borrow the architecture from VHRED [241] for topic discovery, with the key modification that both the encoder RNN and decoder RNN are replaced by layers of a pre-trained language model. Furthermore, we use a transformer multi-head attention in replacement of the LSTM to model the dependence between the latent topic vectors. Unlike VHRED, we are interested in the encoder part to extract the posterior of the latent topic z, rather than the recurrent prior of z in the decoder part since the latter is intended for dialogue generation. We assume each utterance corresponds to a latent variable compacting its internal topic, and we impose sequential dependence on the topic transitions. Figure 5.2 gives an overview of the VAE-based model which aims at learning the latent topic vector during the fine-tuning of the language model.

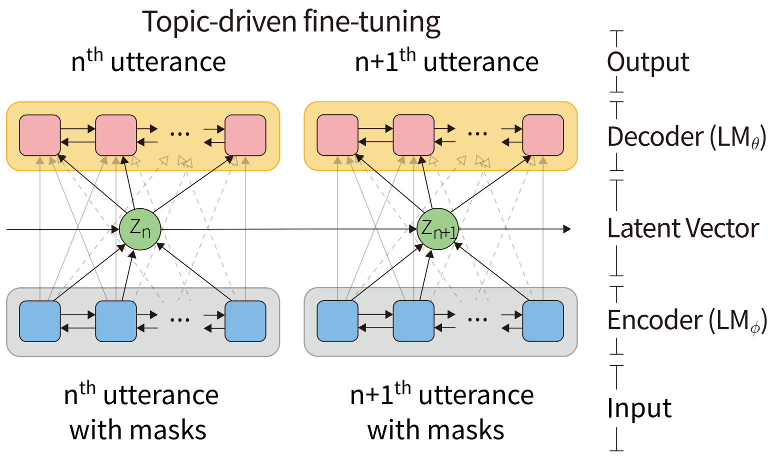


Figure 5.2: Topic-driven fine-tuning of a pre-trained LM.

Specifically, the pre-trained language model is decomposed into two parts, the encoder and the decoder. By retaining the pre-trained weights, we transfer representations from high-resource tasks to the low-resource setting, which is the case for dialogue emotion datasets.

Encoder

The training of topic discovery part of TodKat comprises a VAE at each time step, however with its latent variable dependent on the previous latent code. Each utterance is input to the VAE encoder with a recurrent hidden state, the output of which is a latent vector ideally compressing the topic discussed in the utterance. The latent vectors are tied through a recurrent hidden state to reflect the constraint that they are within the same dialogue. We use LM$_\phi$ to denote the network of lower layers of the language model (before the topic layer) and x_n^L to denote the output from LM_ϕ given the input x_n. The variational distribution for the approximation of the posterior will be

$$q_\phi(z_n|\boldsymbol{x}_{\leq n}, \boldsymbol{z}_{<n}) = \mathcal{N}(z_n|f_{\mu_\phi}(x_n^L, h_{n-1}), f_{\sigma_\phi}(x_n^L, h_{n-1})), \tag{5.3.2}$$

$$\text{where } h_{n-1} = f_\tau(z_{n-1}, x_{n-1}^L), \text{for } n > 1. \tag{5.3.3}$$

Here, $f_{\mu_\phi}(\cdot)$ and $f_{\sigma_\phi}(\cdot)$ are multi-layer perceptrons (MLPs), f_τ can be any transition function (e.g., a recurrent unit). We employ the transformer multi-head attention with its query being the previous latent variable z_{n-1}, that is,

$$f_\tau(z_{n-1}, x_{n-1}^L) = \mathrm{Attention}(z_{n-1}, x_{n-1}^L, x_{n-1}^L). \tag{5.3.4}$$

We initialise $h_0 = \mathbf{0}$ and model the transition between h_{n-1} and h_n by first generating z_n from h_{n-1} using Eq. 5.3.2, then calculating h_n by Eq. 5.3.3.

Decoder

The decoder network reconstructs x_n from z_n at each time step. We use Gaussian distribution for both the generative prior and the variational distribution. Since we want z_n to be dependent on z_{n-1}, the prior for z_n given the preceding hidden state is $p(z_n|h_{n-1}) = \mathcal{N}(z_n|f_{\mu_\gamma}(h_{n-1}), f_{\sigma_\gamma}(h_{n-1}))$. where $f_{\mu_\gamma}(\cdot)$ and $f_{\sigma_\gamma}(\cdot)$ are MLPs. The posterior for z_n is $p_\theta(z_n|\boldsymbol{x}_{\leq n}, \boldsymbol{z}_{<n})$, which is intractable and is approximated by $q_\phi(z_n|\boldsymbol{x}_{\leq n}, \boldsymbol{z}_{<n})$ of Eq. 5.3.2. We denote the higher layers of the language model as LM_θ. Then the reconstruction of $\hat{x}_n$ given z_n and x_n^L can be expressed as:

$$\hat{x}_n = \mathrm{LM}_\theta(z_n, x_n^L). \tag{5.3.5}$$

Note that this is different from dialogue generation in which an utterance is generated from the latent topic vector. Here, we aim to extract the latent topic from the current utterance and therefore train the model to reconstruct the input utterance as specified in Eq. 5.3.5. To make the combination of z_n and x_n^L compatible for LM_θ, we need to perform the latent vector injection. As in [140], we employ the "`Memory`" scheme that z_n becomes an additional input for LM_θ, that is, the input to the higher layers becomes $[z_n, x_n^L]$.

Training

The training objective is the Evidence Lower Bound (ELBO):

$$\mathbb{E}_{q_\phi(\boldsymbol{z}_{\leq N}|\boldsymbol{x}_{\leq N})}[\log p_\theta(\boldsymbol{x}_{\leq N}|\boldsymbol{z}_{\leq N})] - \mathrm{KL}[q_\phi(\boldsymbol{z}_{\leq N}|\boldsymbol{x}_{\leq N})||p(\boldsymbol{z}_{\leq N})] \tag{5.3.6}$$

Eq. 5.3.6 factorizes and the expectation term becomes

$$\mathbb{E}_{q_\phi(\boldsymbol{z}_{\leq N}|\boldsymbol{x}_{\leq N})}\left[\sum_{n=1}^{N} \log p_\theta(x_n|\boldsymbol{z}_{\leq n}, \boldsymbol{x}_{<n})\right], \tag{5.3.7}$$

and the KL term becomes

$$\sum_{n=1}^{N} \mathrm{KL}[q_\phi(z_n|\boldsymbol{x}_{\leq n}, \boldsymbol{z}_{<n})||p(z_n|\boldsymbol{z}_{<n}, \boldsymbol{x}_{<n})], \tag{5.3.8}$$

where $p(z_n|\boldsymbol{z}_{<n}, \boldsymbol{x}_{<n})$ is the prior for z_n. After training, we can extract the topic representation from the encoder part of the model, which is denoted as $z_n = \mathrm{LM}_\phi^{\mathrm{enc}}(x_n)$. Meanwhile, the entire language model has been fine-tuned, which is denoted as $u_n = \mathrm{LM}^{\mathrm{CLS}}(x_n)$.

5.3.3 Knowledge-Aware Transformer

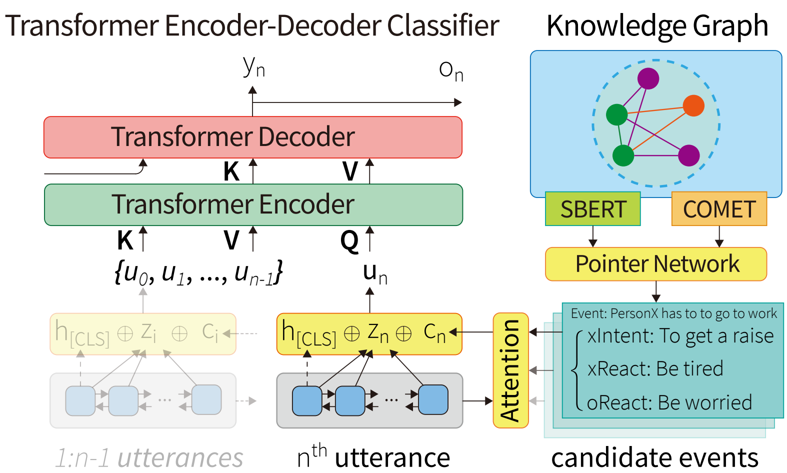


Figure 5.3: Knowledge-aware transformer.

The topic-driven LM fine-tuning stage allows the LM to discover a topic representation from a given utterance. After fine-tuning, we attach the fine-tuned components to a classifier and train the classifier to predict the emotion labels.

We propose to use the Transformer Encoder-Decoder structure as the classifier, and consider the incorporation of commonsense knowledge retrieved from external knowledge sources. In what follows, we first describe how to retrieve commonsense knowledge from a knowledge source, then we present the detailed structure of the classifier.

Commonsense Knowledge Retrieval

We use ATOMIC[2] as a source of external knowledge. In ATOMIC, each node is a phrase describing an event. Edges are triples such as ⟨`event, relation type, event`⟩, linking from one event to another. There are a total of nine relation types, of which three are used: `xIntent`, the intention of the subject (e.g., '*to get a raise*'), `xReact`, the reaction of the subject (e.g., '*be tired*'), and `oReact`, the reaction of the object (e.g., '*be worried*'), since they are defined as the mental states of an event [237].

Given an utterance x_n, we can compare it with every node in the knowledge graph, and retrieve the most similar one. The method for computing the similarity between an utterance and events is SBERT [225]. We extract the top-K events, and obtain their intentions and reactions, which are denoted as $\{e^{sI}_{n,k}, e^{sR}_{n,k}, e^{oR}_{n,k}\}, k = 1, \ldots, K$.

On the other hand, there is a knowledge generation model, called COMET[3], which is trained on ATOMIC. It can take x_n as input and generate the knowledge (e.g., the *intention* or the *reaction*) with the desired event relation types specified (e.g., `xIntent`, `xReact` or `oReact`). The generated knowledge can be unseen in ATOMIC since COMET is essentially a fine-tuned language model. Again, we ask COMET to generate the K most likely events, each with respect to the three event relation types. The produced events are denoted as $\{g^{sI}_{n,k}, g^{sR}_{n,k}, g^{oR}_{n,k}\}$, $k = 1, \ldots, K$.

Knowledge Selection

With the knowledge retrieved from ATOMIC, we build a pointer network [281] to exclusively choose the commonsense knowledge either from SBERT or COMET in order to circumvent the case that no matched events are found by SBERT. The pointer network calculates the probability of choosing the candidate knowledge source as

$$P(\mathbb{I}(x_n, \boldsymbol{e}_n, \boldsymbol{g}_n) = 1) = \sigma([x_n, \boldsymbol{e}_n, \boldsymbol{g}_n]\mathbf{W}_\sigma),$$

[2]`https://homes.cs.washington.edu/~msap/atomic/`

[3]`https://github.com/atcbosselut/comet-commonsense`

where $\mathbb{I}(x_n, \boldsymbol{e}_n, \boldsymbol{g}_n)$ is an indicator function with value 1 or 0, and $\sigma(x) = 1/(1 + \exp(-x))$. We envelope σ with Gumbel Softmax [112] to make the one-hot distribution[4]. The integrated commonsense knowledge is expressed as

$$\boldsymbol{c}_n = \mathbb{I}(x_n, \boldsymbol{e}_n, \boldsymbol{g}_n)\boldsymbol{e}_n + (1 - \mathbb{I}(x_n, \boldsymbol{e}_n, \boldsymbol{g}_n))\boldsymbol{g}_n,$$

where $\boldsymbol{c}_n = \{c^{sI}_{n,k}, c^{sR}_{n,k}, c^{oR}_{n,k}\}_{k=1}^{K}$.

With the knowledge source selected, we proceed to select the most informative knowledge. We design an attention mechanism [11] to integrate the candidate knowledge. Recall that we have a fine-tuned language model which can calculate both the `[CLS]` and topic representations. Here we apply the language model to the retrieved or generated knowledge to obtain the `[CLS]` and the topic representation, denoted as $[\boldsymbol{c}_{n,k}, z_{n,k}]$. The attention mechanism is performed by calculating the dot product between the utterance and every other normalised knowledge tuple:

$$v_k = \tanh([\boldsymbol{c}_{n,k}, z_{n,k}]\mathbf{W}_\alpha), \tag{5.3.9}$$

$$\alpha_k = \frac{\exp(v_k[z_n, u_n]^\top)}{\sum_k \exp(v_k[z_n, u_n]^\top)}, \; \boldsymbol{c}_n = \sum_{k=1}^{K} \alpha_k \boldsymbol{c}_{n,k}.$$

Here, we abuse $\boldsymbol{c}_n$ to represent the knowledge phrases aggregated by k. We further aggregate the $\boldsymbol{c}_n$ by event type using a self-attention and the final event representation is denoted as c_n.

Transformer Encoder-Decoder

We use a Transformer encoder-decoder to map an utterance sequence to an emotion label sequence, thus allowing for modelling the transitional patterns between emotions and taking into account the historical utterances as well. Each utterance is converted to the `[CLS]` representation concatenated with the topic representation z_n and knowledge representation c_n. We enforce a masking scheme in the self-attention sub-layer of the encoder to make the classifier predict emotions in an auto-regressive way, that is, only the past utterances are visible to the encoder. This masking, preventing the query from attending to future keys, is more natural due to the fact that the subsequent utterances are unseen when predicting an emotion of the current utterance. As for the decoder, the output of the previous decoder block is input as a query to the self-attention sub-layer. The training loss for the classifier is the

[4]We have also experimented with a soft gating mechanism by aggregating knowledge from SBERT and COMET in a weighted manner. But the results are consistently worse than those using a hard gating mechanism.

negative log-likelihood expressed as:

$$\mathcal{L} = -\sum_{n=1}^{N} \log\, p_\theta(y_n | \boldsymbol{u}_{\leq n}, \boldsymbol{y}_{<n}),$$

where θ denotes the trainable parameters.

5.4 Experimental Setup

In this section, we present the details of the datasets used, the methods for comparison, and the implementation details of our models.

Datasets We use the following datasets for experimental evaluation:
DailyDialog [148] is collected from daily communications. It takes the Ekman's six emotion types [65] as the annotation protocol, that is, it annotates an utterance with one of the six basic emotions: *anger, disgust, fear, happiness, sadness*, or *surprise*. Those showing ambiguous emotions are annotated as *neutral*.
MELD [217] is constructed from scripts of '*Friends*', a TV series on urban life. Same as DailyDialog, the emotion label falls into Ekman's six emotion types, or *neutral*.
IEMOCAP [36] is built with subtitles from improvised videos. Its emotion labels are *happy, sad, neutral, angry, excited* and *frustrated*.
EmoryNLP [309][5] is also built with conversations from '*Friends*' TV series, but with a slightly different annotation scheme that *disgust, anger* and *surprise* become *peaceful, mad* and *powerful*.

Following Zhong et al. [328] and Ghosal et al. [76], the '*neutral*' label of DailyDialog is not counted in the evaluation due to the extreme imbalance. For MELD and EmoryNLP, we consider a dialogue as a sequence of utterances from the same scene id. Table 5.1 summarizes the statistics of each dataset.

Baselines We compare the performance of TodKat with the following methods:
HiGRU [113] simply inherits the recurrent attention framework that an attention layer is placed between two GRUs to aggregate the signals from the encoder GRU and pass them to the decoder GRU.
DialogueGCN [75] creates a graph from interactions of speakers to take into account the dialogue structure. A Graph Convolutional Network (GCN) is employed to encode the speakers. Emotion labels are predicted with the combinations of the global context and speakers' status.

[5] `https://github.com/emorynlp/emotion-detection`

	DD	MELD	IEMOCAP	EmoryNLP
#Dial.	13,118	1,432	151	827
Train	11,118	1,038	100	659
Dev.	1,000	114	20	89
Test	1,000	280	31	79
#Utt.	102,979	13,708	7,333	9,489
Train	87,170	9,989	4,810	7,551
Dev.	8,069	1,109	1,000	954
Test	7,740	2,610	1,523	984
#Cat.	7	7	6	7

Table 5.1: Statistics of the benchmarks for dialogue emotion detection. Every benchmark has provided a training set, a development set and a testing set, which is detailed in the number of utterances.

KET [328] is the first model which integrates common-sense knowledge extracted from ConceptNet and emotion information from an emotion lexicon into conversational text. A Transformer Encoder is employed to handle the influence of past utterances.
COSMIC [76] is the state-of-the-art approach that leverages ATOMIC for improved emotion detection. COMET is employed in their model to retrieve the event-eccentric commonsense knowledge phrases from ATOMIC.

Settings We modified the script[6] of language model fine-tuning in the Hugging Face library [294] for the implementation of topic-driven fine-tuning. We use one transformer encoder layer. As for the decoder, there are N layers where N is the number of utterances in a dialogue. On each training set, we train the topic model for 3 epochs, with learning rate set to 5e-5 to prevent overfitting to the low-resource dataset. The classifier is built on the Transformers[7] package in Hugging Face. The language model we employ is RoBERTa [155]. Each utterance is padded by the `<pad>` token of RoBERTa if it is less than the maximum length of 128. The maximum number of utterances in a dialogue is set to 36, 25, 72 and 25 respectively for DD [148] [8], MELD [217] [9], IEMOCAP [36] [10] and EmoryNLP [309] [11]. Dialogues with shorter lengths are padded with `NULL`. It is worth noting that this step is

[6] `https://huggingface.co/transformers/v2.0.0/examples.html`
[7] `https://huggingface.co/transformers/`
[8] `http://yanran.li/dailydialog.html`
[9] `https://github.com/declare-lab/MELD`
[10] `https://sail.usc.edu/iemocap/iemocap_release.htm`
[11] `https://github.com/emorynlp/emotion-detection`

performed after RoBERTa due to the random noises introduced by RoBERTa. The number of retrieved or generated events from ATOMIC under the relation types '*intentions*' and '*reactions* is set to 5, respectively, i.e., $K = 5$.

5.5 Results and Analysis

Comparison with Baselines Experiment results of TODKAT and its ablations are reported in Table 5.2. HiGRU and DialogueGCN results were produced by running the code published by the authors on the four datasets.

Among the baselines, COSMIC gives the best results. Our proposed TODKAT outperforms COSMIC on both MELD and EmoryNLP in weighted Avg-F1 and Micro-F1. TODKAT also achieves superior results than COSMIC on DailyDialogue in Macro-F1 and gives nearly the same result in Micro-F1. TODKAT is inferior to COSMIC on IEMOCAP. It is however worth mentioning that COSMIC was trained with 132 instances on this dataset, while for all the other models the training-and-validation split is 100 and 20. As such, the IEMOCAP results reported on COSMIC [76] are not directly comparable here. COSMIC also incorporates the commonsense knowledge from ATOMIC but with the modified GRUs. Our proposed TODKAT, built upon the topic-driven Transformer, appears to be a more effective architecture for dialogue emotion detection. Compared with KET, the improvements are much more significant, with over 7% increase on MELD, and close to 5% gain on DailyDialog. KET is also built on the Transformer, but it considers each utterance in isolation and applies commonsense knowledge from ConceptNet. TODKAT, on the contrary, takes into account the dependency of previous utterances and their associated emotion labels for the prediction of the emotion label of the current utterance. DialogueGCN models interactions of speakers and it performs slightly better than KET. But it is significantly worse than TODKAT. It seems that topics might be more useful in capturing the dialogue context.

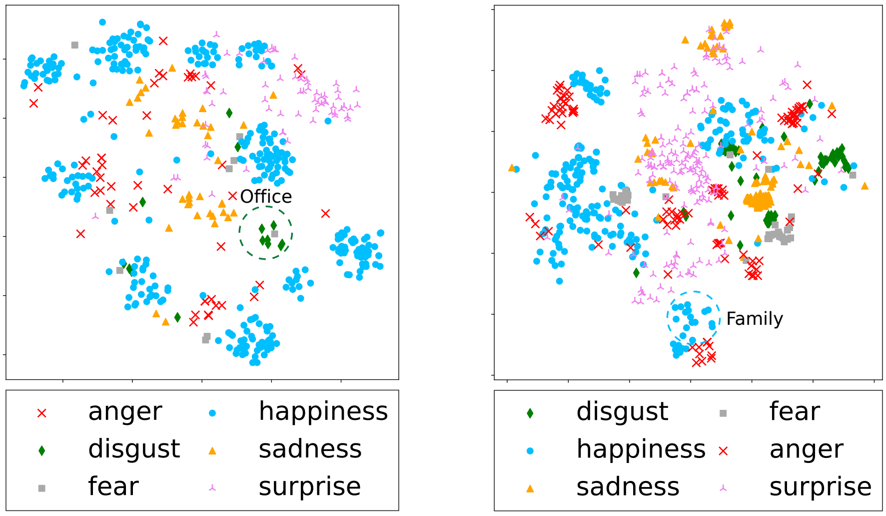
Ablation Study The lower half of Table 5.2 presents the F1 scores with the removal of various components from TODKAT. It can be observed that with the removal of the topic component, the performance of TODKAT drops consistently across all datasets except IEMOCAP in which we observe a slight increase in both weighted average F1 and Micro-F1. This might be attributed to the size of the data since IEMOCAP is the smallest dataset evaluated here, and the small dataset size doesn't favour the discovery of topics. Without using the commonsense knowledge ('−KB'), we observe a more drastic performance drop compared to all other com-

Models	DailyDialog		MELD		IEMOCAP		EmoryNLP	
	Macro-F1 - neutral	Micro-F1 - neutral	weighted Avg-F1	Micro-F1	weighted Avg-F1	Micro-F1	weighted Avg-F1	Micro-F1
HiGRU	0.4904	0.5190	0.5681	0.5452	0.5854	0.5828	0.3448	0.3354
DialogueGCN	0.4995	0.5373	0.5837	0.5617	0.6085	0.6063	0.3429	0.3313
KET	–	0.5348	0.5818	–	0.5956	–	0.3439	–
COSMIC	0.5105	**0.5848**	0.6521	–	**0.6528**[1]	–	0.3811	–
TodKat	**0.5256**	0.5847	**0.6547**	**0.6724**	0.6133	0.6111	**0.3869**	**0.4238**
−Topics	0.5136	0.5549	0.6408	0.6586	0.6281	**0.6260**	0.3747	0.4014
−KB	0.5003	0.5344	0.6152	0.6318	0.5896	0.5738	0.3460	0.3806
$\text{Kat}_{\text{SBERT}}$	0.5173	0.5578	0.6239	0.6387	0.6097	0.6069	0.3622	0.3913
$\text{Kat}_{\text{Comet}}$	0.5102	0.5462	0.6379	0.6487	0.6277	0.6254	0.3714	0.3984

Table 5.2: The F1 results of the dialogue emotion detectors on four benchmarks. Here we denote the proposed model as TodKat, of which the results are an average of ten runs. The ablations of different components are reported separately in the bottom, where the model without the incorporation of latent topics is denoted as '−Topics', transformer encoder-decoder structure without the use of a knowledge base is denoted as '−KB'. $\text{Kat}_{\text{Comet}}$ and $\text{Kat}_{\text{SBERT}}$ uses the commonsense knowledge obtained with Comet and SBERT, respectively. Results of KET and COSMIC are from [328] and [76], respectively.

[1] This is incomparable due to the larger training set.

ponents, with a nearly 2.2% drop in F1 on EmoryNLP, showing the importance of employing commonsense knowledge for dialogue emotion detection. Comparing two different ways of extracting knowledge from ATOMIC, direct retrieval using SBERT or generation using COMET, we observe mixed results. Overall, the Transformer Encoder-Decoder with a pointer network is a conciliator between the two methods, yielding a balanced performance across the datasets.



(a) DailyDialog (b) MELD

Topic	Utterances	Emotion
Office	A: How are you doing, Christopher? B: To be honest, I'm really fed up with work at the moment. I need a break! A: Are you doing anything this weekend? B: I have to work on Saturday all day! I really hate my job!	disgust
Family	A: Yeah, I-I heard. I think it's great! Ohh, I'm so happy for you! B: I can't believe you're getting married! C: Yeah. D: Monica and Rachel made out.	happy

(c) Representative utterances and their topics

Figure 5.4: T-SNE visualization on DailyDialog and MELD. Utterances with the same colour have the same emotion label as shown in the last column. Visualization and highlight of the neutral utterances are omitted for clarity. Each cluster is exemplified by a group of utterances.

Relationships between Topics and Emotions To investigate the effectiveness of the learned topic vectors, we perform t-SNE [276] on the test set to study the relationship between the learned topic vectors and the ground-truth emotion labels. The results on DailyDialog and MELD are illustrated in Figure 5.4(a) and (b). Latent topic vectors of utterance are used to plot the data points, whose colors indicate their ground-truth emotion labels. We can see that the majority of the topic vectors cluster into polarized groups. Few clusters are bearing a mixture of polarity, possibly due to the background topics such as greetings in the datasets.

Topics can be interpreted using the attention scores of Eq. 5.3.4. The top-10 most-attended words are selected as the representative words for each utterance. As in [57], we construct bag-of-words[12] that represent 141 distinct topics. Given the attended words of an utterance cluster are grouped based on their latent topic representations, we label the word collection with the dominant theme name. We refer to the theme names as topics in Figure 5.4c. It can be observed that utterances associated with office carry *'disgust'* emotions, while those related to *family* are prone to be *'happy'*. There are also cases where similar utterances exhibit different emotions due to the changes in topics, e.g., "A: Johnny died yesterday, we knew that it was coming, but. B: Like just last week, he was doing so well." and "A: Then all of a sudden they give him a microphone, he asked me to marry him, like, onstage. B: He scored points.", showing that the emotion of interlocutors heavily depends on the topics they are talking about. Usually, topics play a major part in determining the emotion, but emotional transition also contributes to the changes.

We further compute the Spearman's rank-order correlation coefficient to quantitatively verify the relationship between the topic vectors and emotion vectors. For an utterance pair, a similarity score is obtained separately for their corresponding topic vectors as well as their emotion vectors. We then sort the list of emotion vector pairs according to their similarity scores to see how its order coordinates with that of topic vector pairs using the Spearman's rank-order correlation coefficient. The results are $0.60, 0.58, 0.42$ and 0.54 with p-values $\ll 0.01$ respectively for DailyDialog, MELD, IEMOCAP and EmoryNLP, showing that there is a strong correlation between the clustering of topics and that of emotion labels. IEMOCAP has the lowest correlation score, which is inline with the results in Table 2 that the discovered latent topics did not improve the emotion classification results.

[12] Word lists and their corresponding theme names are crawled from `https://www.enchantedlearning.com/wordlist/`.

Dataset	Relation Type	
	$\{sI, sR, oR, sE, oE\}$	All
DailyDialog	0.5718↓	0.5664↓
MELD	0.6578↓	0.6460↓
IEMOCAP	0.6163↑	0.6073↓
EmoryNLP	0.4055↓	0.3892↓

Table 5.3: Micro-F1 scores of TodKat with more commonsense relation types retrieved from Atomic included for training. Here, "*sE*" and "*oE*" represent *effect of subject* and *effect of object*, respectively. "All" denotes the incorporation of all nine commonsense relation types from Atomic.

Impact of Relation Type We investigate the impact of commonsense relation types on the performance of TodKat. We expand the relation set to five relation types and all nine relation types, respectively. According to [237], there are other relation types including {*sNeed*, *sWant*, *oWant*, *sEffect*, *oEffect*}, which identifies the prerequisites and post conditions of the given event, and {*sAttr*}, the "If-Event-Then-Persona" category of relation type that describes how the subject is perceived by others. We calculate the Micro-F1 scores of TodKat with these two categories of relation types added step by step. From Table 5.3 we can conclude that the inclusion of two extra relation types or all relation types degrades the F1 scores on almost all datasets. An exception occurs on IEMOCAP where the F1 score rises by 0.5% when adding "*sE*" and "*oE*" relations, possibly due to the fact that the dataset is abundant in events. Hence the extra event descriptions offer complementary knowledge to some extent. While on other datasets neither the incorporation of "If-Event-Then-Event" nor the incorporation of "If-Event-Then-Persona" relation types could bring any benefit.

Impact of Attention Mechanism With the knowledge retrieved from Atomic or generated from Comet, we are able to infer the possible intentions and reactions of the interlocutors. However, not all knowledge phrases contribute the same to the emotion of the focused utterance. We study the attention mechanism in terms of selecting the relevant knowledge. We show in Table 5.4 a heat map of the attention scores in Eq. 5.3.9 to illustrate how the topic-driven attention could identify the most salient phrase. The utterance '*Oh my God, you're a freak.*' will be erroneously categorized as '*mad*' without using the topic-driven attention (shown in the last row of Table 5.4).

Dialogue Context	A: Alright, go on.	Neutral
	B: Ok, I have to sleep on the west side because I grew up in California and otherwise the ocean would be on the wrong side.	Neutral
	A: Oh my God, you're a freak.	Joyful
	B: Yeah. How about that.	Neutral
Topic-Driven Attention	A wants to be liked	
	A wants to be accepted	
	A wants to be a freak	
	A will feel satisfied	
	A will feel ashamed	Joyful ✓
	A will feel happy	
	B will feel impressed	
	B will feel disgusted	
	B will feel surprised	
	A: Oh my God, you're a freak.	Mad ✗

Table 5.4: Illustration of Attention mechanism in Eq. 5.3.9 that helps distinguish the retrieved knowledge.

5.6 Summary

We have designed a Topic-Driven and Knowledge-Aware Transformer model that incorporates topic representation and the commonsense knowledge from ATOMIC for emotion detection in dialogues. A topic-augmented language model based on fine-tuning has been developed for topic extraction. Pointer networks and additive attention have been explored for knowledge selection. All novel components have been integrated into the Transformer Encoder-Decoder structure, enabling Seq2Seq prediction. Empirical results show the model's effectiveness in topic representation learning and knowledge integration, which have both boosted the performance of emotion detection.

Chapter 6

Disentangled Learning of Stance and Aspect Topics for Attitude Detection

Chapter Abstract

We target the disentanglement of tweets regarding stance and aspect topics in this chapter for the benefit of vaccination attitude detection. Our goal is to detect the stance expressed in a tweet (i.e., '*pro-vaccination*', '*anti-vaccination*', or '*neutral*'), identify a text span that indicates the concerning aspect of vaccination, and cluster tweets into groups that share similar aspects. To this end, we propose a novel latent representation learning model that jointly learns a stance classifier and disentangles the latent variables capturing stance and aspect. The model employs a semi-supervised framework that comprises an LM-based VAE and a fine-tuned text span predictor. We build a dataset called VADet, on which we validate the proposed approach. The results show that the VADet model is able to learn disentangled stance and aspect topics, and outperforms several aspect-based sentiment analysis models on both stance detection and tweet clustering.

6.1 Introduction

The aim of vaccine attitude detection in social media is to extract people's opinions towards vaccines by analysing their online posts. This is closely related to aspect-based sentiment analysis in which both aspects and related sentiments need to be identified. Previous research has been largely focused on product reviews and relied on aspect-level sentiment annotations to train models [16], where aspect-opinions are extracted as triples [208], polarized targets [166] or sentiment spans [92]. However, for the task of vaccine attitude detection on Twitter, such a volume of annotated data is barely available [130, 206]. This scarcity of data is compounded by the diversity of attitudes, making it difficult for models to identify all aspects discussed in posts [189].

As representative examples, consider the two tweets about personal experiences with vaccination at the top of Figure 6.1. The two tweets, despite addressing a common aspect (vaccine side-effects), express opposite stances towards vaccines. However, the aspect and the stances are so fused together that the whole of the tweets need to be considered to derive the proper labels, making it difficult to disentangle them using existing methodologies. Additionally, in the case of vaccines attitude analysis, there is a wide variety of possible aspects discussed in posts, as shown at the bottom of Figure 6.1, where one tweet ironically addressed vaccine side-effects and the second one expressed instead of specific political concerns. This is different from traditional aspect-based sentiment analysis on product reviews where only a small number of aspects need to be pre-defined.

The recently developed framework for integrating Variational Auto-Encoder (VAE) [123] and Independent Component Analysis (ICA) [121] sheds light on this problem. VAE is an unsupervised method that can be used to glean information that must be retained from the vaccine-related corpus. Meanwhile, a handful of annotations would induce the separation of independent factors following the ICA requirement for prior knowledge and inductive biases [108, 158, 159]. To this end, we could disentangle the latent factors that are either specific to the aspect or to the stance, and improve the quality of the latent semantics learned from unannotated data.

We frame the problem of vaccine attitude detection as a joint aspect span detection and stance classification task, assuming that a tweet, which is limited to 280 characters, would usually only discuss one aspect. In particular, we extend a pre-trained language model (LM) by adding a topic layer, which aims to model the topical theme discussed in a tweet. In the absence of annotated data, the

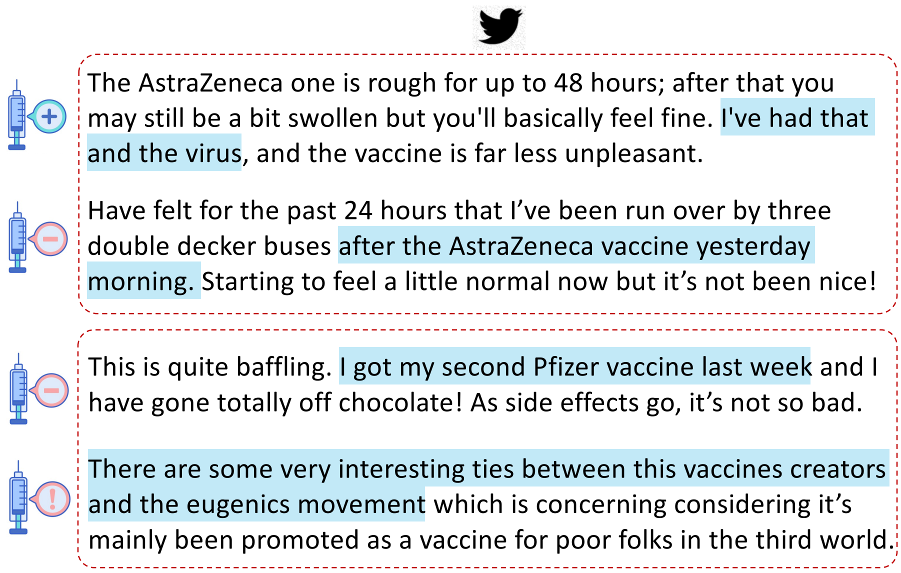


Figure 6.1: **Top:** Expressions of aspects entangled with expressions of opinions. **Bottom:** Vaccine attitudes can be expressed towards a wide range of aspects/topics relating to vaccination, making it difficult to pre-define a set of aspect labels as opposed to corpora typically used for aspect-based sentiment analysis.

topic layer is trained to reconstruct the input message built on VAE. Given the annotated data, where each tweet is annotated with an aspect span and a stance label, the learned topic can be disentangled into a stance topic and an aspect topic. The stance topic is used to predict the stance label of the given tweet, while the aspect topic is used to predict the start and ending positions of the aspect span. By doing so, we can effectively leverage both unannotated and annotated data for model training. To evaluate the effectiveness of our proposed model for vaccine attitude detection on Twitter, we have collected over 1.9 million tweets relating to COVID vaccines between February and April 2021. We have further annotated 2,800 tweets with both aspect spans and stance labels. In addition, we have also used an existing Vaccination Corpus[1] in which 294 documents related to the online vaccination debate have been annotated with opinions towards vaccination. Our experimental results on both datasets show that the proposed model outperforms existing opinion triple extraction model and BERT QA model on both aspect span extraction and stance classification. Moreover, the learned latent aspect topics allow the clustering of user attitudes towards vaccines, facilitating easier discovery of positive and negative attitudes in social media. The contribution of this work can

[1] https://github.com/cltl/VaccinationCorpus

be summarised as follows[2]:

- We have proposed a novel semi-supervised approach for joint latent stance/aspect representation learning and aspect span detection;
- The developed disentangled representation learning facilitates better attitude detection and clustering;
- We have constructed an annotated dataset for vaccine attitude detection.

6.2 Related Work

This work is related to three lines of research: aspect-based sentiment analysis, disentangled representation learning, and vaccine attitude detection.

Aspect-Based Sentiment Analysis (ABSA) aims to identify the aspect terms and their polarities from text. Much work has been focusing on this task. The techniques used include Conditional Random Fields (CRFs) [171], Bidirectional Long Short-Term Memory networks (BiLSTMs) [17], Convolutional Neural Networks (CNNs) [321], Attention Networks [211, 305], DenseLSTMs [295], NestedLSTMs [186], Graph Neural Networks [311] and their combinations [283, 286], to name a few. Zhang et al. [317] framed this task as text span detection, where they used text spans to denote aspects. The same annotation scheme was employed in [145], where intra-word attentions were designed to enrich the representations of aspects and predict their polarities. Li et al. [146] formalized the task as a sequence labeling problem under a unified tagging scheme. Their follow-up work [147] explored BERT for end-to-end ABSA. Peng et al. [208] modified this task by introducing opinion terms to shape the polarity. A similar modification was made in [325] to extract aspect-opinion pairs. Position-aware tagging was introduced to entrench the offset between the aspect span and opinion term [302]. More recently, instead of using pipeline approaches or sequence tagging, Barnes et al. [16] adapted syntactic dependency parsing to perform aspect and opinion expression extraction, and polarity classification, thus formalizing the task as structured sentiment analysis.

Disentangled representation learning Deep generative models learn the hidden semantics of text, of which many attempt to capture the independent latent factor to steer the generation of text in the context of NLP [63, 103, 115, 140, 143, 210]. The majority of the aforementioned work employs VAE [125] to learn controllable

[2]Our source code and dataset are available at `http://github.com/somethingx1202/VADet`

factors, leading to the abundance of VAE-based models in disentangled representation learning [35, 43, 94]. However, previous studies show that unsupervised learning of disentanglement by optimising the marginal likelihood in a generative model is impossible [157]. While it is also the case that non-linear ICA is unable to uncover the true independent factors, Khemakhem et al. [121] established a connection between those two strands of work, which is of particular interest to us since the proposed framework learns to approximate the true factorial prior given few examples, recovering a disentangled latent variable distribution on top of additionally observed variables. In the proposed approach, stance labels and aspect spans are additionally observed on a handful of data, which could be used as inductive biases that make disentanglement possible.

Vaccine attitude detection Very little literature exists on attitude detection for vaccination. In contrast, there is growing interest in Covid-19 corpus construction [250]. Of particular interest to us, Banda et al. [13] built an on-going tweet dataset that traces the development of Covid-19 by 3 keywords: "coronavirus", "2019nCoV" and "corona virus". Hussain et al. [105] utilized hydrated tweets from the aforementioned corpus to analyze the sentiment towards vaccination. They used lexicon-based methods (i.e., VADER and TextBlob) and pre-trained BERT to classify the sentiment in order to gain insights into the temporal sentiment trends. A similar approach has been proposed in [102]. Lyu et al. [165] employed a topic model to discover vaccine-related themes in twitter discussions and performed sentiment classification using lexicon-based methods. However, none of the work above constructed datasets about vaccine attitudes, nor did they train models to detect attitudes. Morante et al. [189] built the Vaccination Corpus (VC) with events, attributions and opinions annotated in the form of text spans, which is the only dataset available to us to perform attitude detection.

6.3 Proposed Approach

The goal of our work is to detect the stance expressed in a tweet (i.e., '*pro-vaccination*', '*anti-vaccination*', or '*neutral*'), identify a text span that indicates the concerning aspect of vaccination, and cluster tweets into groups that share similar aspects. To this end, we propose a novel latent representation learning model that jointly learns a stance classifier and disentangles the latent variables capturing stance and aspect respectively. Our proposed Vaccine Attitude Detection (VADet) model is firstly trained on a large amount of unannotated Twitter data to learn latent topics via masked Language Model (LM) learning. It is then fine-tuned on a small amount of Twitter data annotated with stance labels and aspect text spans for simultaneously stance classification and aspect span start/end position detection. The rationale is that the inductive bias imposed by the annotations would encourage the disentanglement of latent stance topics and aspect topics. In what follows, we will present our proposed VADET model, first under the masked LM learning and later extended to the supervised setting for learning disentangled stance and aspect topics.

6.3.1 VADet in the masked LM learning

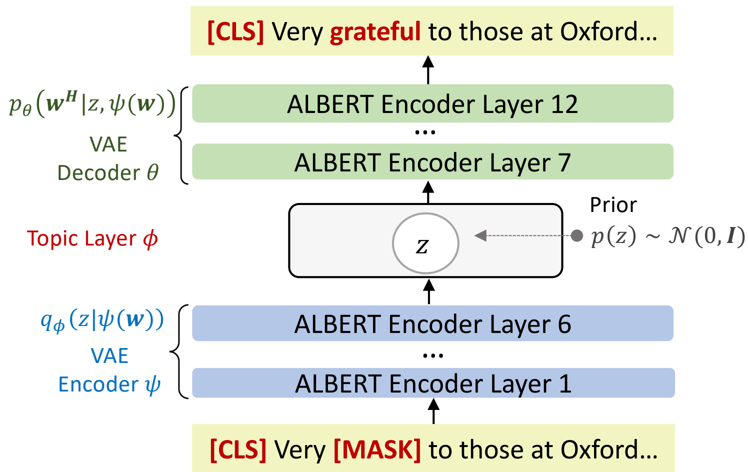


Figure 6.2: VADET in masked language model learning. The latent variables are encoded via the topic layers incorporated into the masked language model.

We insert a topic layer into a pre-trained language model such as ALBERT, as shown in Figure 6.2, allowing the network to leverage pre-trained information while fine-tuned on an in-domain corpus. We assume that there is a continuous latent variable z involved in the language model to reconstruct the original text from the masked tokens. We retain the weights of a language model and learn the latent representation during the fine-tuning. More concretely, the topic layer partitions a language model into lower layers and higher layers denoted as ψ and θ, respectively. The lower layers constitute the Encoder that parameterizes the variational posterior distribution denoted as $q_\phi(z|\psi(\mathbf{w}))$, while the higher layers reconstruct the input tokens, which is referred to as the Decoder.

The objective of VAE is to minimize the KL-divergence between the variational posterior distribution and the approximated posterior. This is equivalent to maximizing the Evidence Lower BOund (ELBO) expressed as:

$$\mathbb{E}_{q_\phi(z|\psi(\mathbf{w}))}[\log p_\theta(\mathbf{w}^H|z,\psi(\mathbf{w}))] - \mathrm{KL}[q_\phi(z|\psi(\mathbf{w}))||p(z)], \tag{6.3.1}$$

where $q_\phi(z|\psi(\mathbf{w}))$ is the encoder and $p_\theta(\mathbf{w}^H|z,\psi(\mathbf{w}))$ is the decoder. Here, $\mathbf{w} = [w_{\texttt{CLS}}, w_{1:n}]$, since the special classification embedding $w_{\texttt{CLS}}$ is automatically prepended to the input sequence [60], $\mathbf{w}^H$ denotes the reconstructed input.

Following [123], we choose a standard Gaussian distribution as the prior, denoted as $p(z)$, and the diagonal Gaussian distribution as the variational distribution, which is analogous to a regularizer [265]. The decoder computes the probability of the original token given the latent variable sampled from the Encoder. We use the Memory Scheme [140] to concatenate z and $\psi(\mathbf{w})$, making the latent representation compatible for higher layers of the language model. Then the latent presentation z is passed to θ to reconstruct the original text.

6.3.2 VADet with disentanglement of aspect and stance

One of the training objectives of vaccine attitude detection is to detect the text span that indicates the aspect and to predict the associated stance label. Existing approaches rely on structured annotations to indicate the boundary and dependency between aspect span and opinion words [16, 302], or use a two-stage pipeline to detect the aspect span and the associated opinion separately [208]. The problem is that the opinion expressed in a tweet and the aspect span often overlap. To mitigate this issue, we instead separate the stance and aspect from their representations in the latent semantic space, that is, disentangling latent topics learned by VADet into latent stance topics and latent aspect topics.

A recent study in disentangled representation learning [157] shows that unsupervised learning of disentangled representations is theoretically impossible from i.i.d. observations without inductive biases, such as grouping information [29] or access to labels [159, 273]. As such, we extend our model to a supervised setting in which the disentanglement of the latent vectors can be trained on annotated data.

Figure 6.3 outlines the overall structure of VADet in the supervised setting. On the right-hand side, we show VADet learned from the annotated aspect text span $[w_a : w_b]$ under masked LM learning. The latent variable z_a encodes the hidden semantics of the aspect expression. We posit that the aspect span is generated from a latent representation with a standard Gaussian distribution being its prior. The

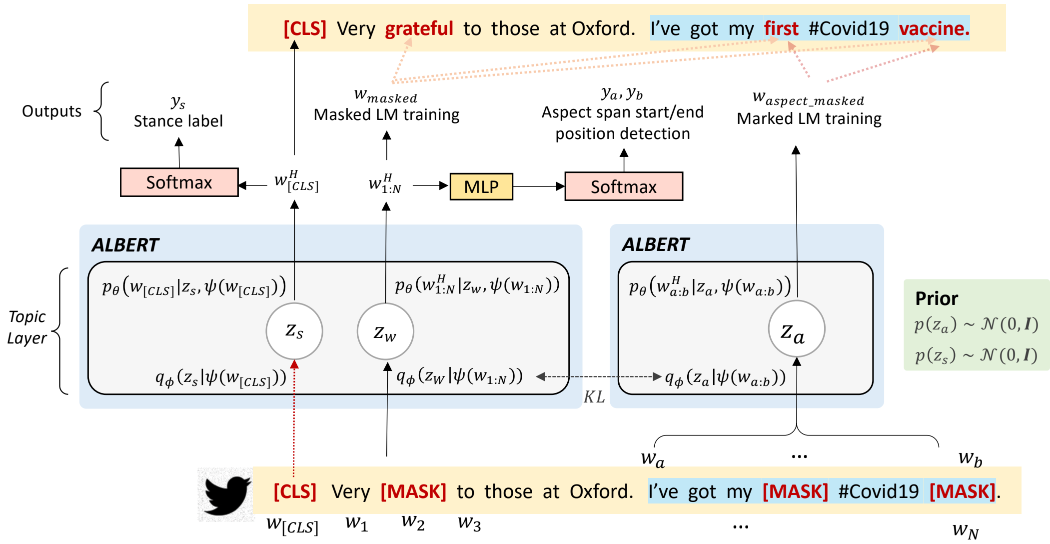


Figure 6.3: VADET in supervised learning. The text segment highlighted in blue is the annotated aspect span. The right part learns latent aspect topic z_a from aspect text span $[w_a : w_b]$ only under masked LM learning. The left part learns jointly latent stance topic z_s and latent aspect topic z_w from the whole input text, and trained simultaneously for stance classification and aspect start/end position detection.

ELBO for reconstructing the aspect text span is:

$$\mathcal{L}_A = \mathbb{E}_{q_\phi(z_a|\psi(w_{a:b}))}[\log p_\theta(w^H_{a:b}|z_a, \psi(w_{a:b}))] - \mathrm{KL}[q_\phi(z_a|\psi(w_{a:b}))||p(z_a)], \quad (6.3.2)$$

where $w^H_{a:b}$ denotes the reconstructed aspect span. Ideally, the latent variable z_a does not encode any stance information and only captures the aspect mentioned in the sentence. Therefore, the z_s for the language model on the right hand side is detached and the reconstruction loss for `[CLS]` is set free.

On the left hand side of Figure 6.3, we train VADET on the whole sentence. The input to VADET is formalized as: '`[CLS] text`'. Instead of mapping an input to a single latent variable z, as in masked LM learning of VADET, the input is now mapped to a latent variable decomposing into two components, $[z_s, z_w]$, one for the stance and another for the aspect. We use a conditionally factorized Gaussian prior over the latent variable $z_w \sim p_\theta(z_w|w_{a:b})$, which enables the separation of z_s and z_w since the diagonal Gaussian is factorized and the conditioning variable $w_{a:b}$ is observed.

We establish an association between z_w and z_a by specifying $p_\theta(z_w|w_{a:b})$ to be the encoder network of $q_\phi(z_a|w_{a:b})$, since we want the latent semantics of aspect span to encourage the disentanglement of attitude in the latent space. In other words, the prior of z_w is configured as the approximate posterior of z_a to enforce the association between the disentangled aspect in a sentence and the *de facto* aspect. As a result, the ELBO for the original text is written as

$$\begin{aligned}\mathbb{E}_{q_\phi(z_w|\psi(\mathbf{w}))}&[\log p_\theta(\mathbf{w}^H|z_w, \psi(\mathbf{w}))] \\ &-\mathrm{KL}[q_\phi(z_w|\psi(\mathbf{w}))||q_\phi(z_w|\psi(w_{a:b}))],\end{aligned} \tag{6.3.3}$$

where $\mathbf{w}^H$ denotes the reconstructed input text, $z_w|\mathbf{w} \sim \mathcal{N}(\mu_\phi(\psi(\mathbf{w})), \sigma^2_\phi(\psi(\mathbf{w})))$. The KL-divergence allows for some variability since there might be some semantic drift from the original semantics when the aspect span is placed in a longer sequence.

The annotation of the stance label provides an additional input. To exploit this inductive bias, we enforce the constraint that z_s participates in the generation of `[CLS]`, which follows an approximate posterior $q_\phi(z_s|\psi(\mathbf{w}_{\texttt{[CLS]}}))$. We place the standard Gaussian as the prior over $z_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and obtain the ELBO

$$\begin{aligned}\mathbb{E}_{q_\phi(z_s|\psi(\mathbf{w}_{\texttt{[CLS]}}))}&[\log p_\theta(w^H_{\texttt{[CLS]}}|z_s, \psi(\mathbf{w}_{\texttt{[CLS]}}))] \\ &-\mathrm{KL}[q_\phi(z_s|\psi(\mathbf{w}_{\texttt{[CLS]}}))||p(z_s)]\end{aligned} \tag{6.3.4}$$

Since the variational family in Eq. 6.3.1 are Gaussian distributions with a diagonal covariance, the joint space of $[z_s, z_w]$ factorizes as $q_\phi(z_s, z_w|\psi(\mathbf{w})) = q_\phi(z_s|\psi(\mathbf{w}))q_\phi(z_w|\psi(\mathbf{w}))$ [191].

Assuming z_w to be solely dependent on $\psi(w_{1:n})$, we obtain the ELBO for the entire input sequence:

$$\begin{aligned}\mathcal{L}_S = \mathbb{E}_{q_\phi(z_w)}\mathbb{E}_{q_\phi(z_s)}&[\log p_\theta(\mathbf{w}^H|z, \psi(\mathbf{w}))] \\ &-\mathrm{KL}\left[q_\phi(z_w|\psi(w_{1:n}))||q_\phi(z_w|\psi(w_{a:b}))\right] \\ &-\mathrm{KL}[q_\phi(z_s|\psi(\mathbf{w}))||p(z_s)].\end{aligned} \tag{6.3.5}$$

Note that the expectation term can be decomposed into the expectation term in Eq. 6.3.3 and Eq. 6.3.4 according to the decoder structure. The derivation is elaborated on below:

Derivation of the Decomposed ELBO Unsupervised training is based on maximizing the Evidence Lower Bound (ELBO):

$$\mathbb{E}_{q_\phi(z_s, z_w|\psi(\mathbf{w}))}[\log p_\theta(\mathbf{w}|z_s, z_w, \psi(\mathbf{w}))] - \mathrm{KL}[q_\phi(z_s, z_w|\psi(\mathbf{w}))||p(z_s, z_w)],$$

where z is partitioned into z_s and z_w. Like standard VAE [123], the variational distribution is a multivariate Gaussian with a diagonal covariance:

$$q_\phi(z_s, z_w|\psi(\mathbf{w})) = \mathcal{N}(z_s, z_w|\mu, \sigma^2 I),$$

where $\mu = [\mu^s, \mu^w]$ and $\sigma = [\sigma^s, \sigma^w]$. Since the coveriance matrix is diagonal, z_s and z_w are uncorrelated. Therefore, the joint probability is decomposed into:

$$q_\phi(z_s, z_w|\psi(\mathbf{w})) = q_\phi(z_s|\psi(\mathbf{w}))q_\phi(z_w|\psi(\mathbf{w})),$$

where $q_\phi(z_s|\psi(\mathbf{w})) = \mathcal{N}(z_s|\mu^s, \sigma^s)$, ϕ are the variational parameters. The prior of $[z_s, z_w] \sim \mathcal{N}(z_s, z_w|\mathbf{0}, I)$ can also be decomposed into the product of $p(z_s)$ and $p(z_w)$, then the KL term becomes:

$$\mathrm{KL}[q_\phi(z_s|\psi(\mathbf{w}))||p(z_s)] + \mathrm{KL}[q_\phi(z_w|\psi(\mathbf{w}))||p(z_w)].$$

As for the decoder $p_\theta(\mathbf{w}|z_s, z_w, \psi(\mathbf{w}))$, the reconstruction of each masked token and $w_{\texttt{[CLS]}}$ are independent from each other, i.e., they are not predicted in an autoregressive way. Therefore, the joint probability is decomposed into:

$$p_\theta(\mathbf{w}|z_s, z_w, \psi(\mathbf{w})) = p_\theta(w_{\texttt{[CLS]}}|z_s, z_w, \psi(\mathbf{w}))\, p_\theta(w_{1:n}|z_s, z_w, \psi(\mathbf{w}))$$

We customize the decoder network to make $w_{\texttt{[CLS]}}$ solely dependent on z_s, and obtain

$$\mathbb{E}_{q_\phi(z_s)}\mathbb{E}_{q_\phi(z_w)}[\log p_\theta(w_{\texttt{[CLS]}}|z_s, \psi(\mathbf{w})) \ + \log p_\theta(w_{1:n}|z_w, \psi(\mathbf{w}))]$$

Here, we omit $\psi(\mathbf{w})$ for notational simplicity. Given the supervision of annotated aspect spans, the prior of z_w is constrained by $q_\phi(z_w|\psi(w_{a:b}))$ (a.k.a., the encoder of $w_{a:b}$), this will change the KL term into:

$$\mathrm{KL}[q_\phi(z_s|\psi(\mathbf{w}))||p(z_s)] + \mathrm{KL}[q_\phi(z_w|\psi(w_{1:n}))||q_\phi(z_w|\psi(w_{a:b}))],$$

and finally the ELBO is expressed as

$$\begin{aligned}
&\mathbb{E}_{q_\phi(z_s)}[\log p_\theta(w_{\texttt{[CLS]}}|z_s, \psi(\mathbf{w}))] \\
&+ \mathbb{E}_{q_\phi(z_w)}[\log p_\theta(w_{1:n}|z_w, \psi(\mathbf{w}))] \\
&- \mathrm{KL}[q_\phi(z_s|\psi(\mathbf{w}))||p(z_s)] \\
&- \mathrm{KL}[q_\phi(z_w|\psi(w_{1:n}))||q_\phi(z_w|\psi(w_{a:b}))].
\end{aligned}$$

Training objective Finally, we perform stance classification and classification for the starting and ending position over the aspect span of a tweet. We use the negative log-likelihood loss for both the stance label and aspect span:

$$\mathcal{L}_s = -\log p(y_s | w^H_{\texttt{[CLS]}}),$$
$$\mathcal{L}_a = -\log p(y_a | \text{MLP}(w^H_{1:n})) - \log p(y_b | \text{MLP}(w^H_{1:n})),$$

where MLP is a fully-connected feed-forward network with tanh activation, y_s is the predicted stance label, y_a and y_b are the starting and ending position of the aspect span. The overall training objective in the supervised setting is:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_a - \mathcal{L}_S - \mathcal{L}_A$$

6.4 Experimental Setup

6.4.1 Datasets

We evaluate our proposed VADet and compare it against baselines on two vaccination attitude datasets.

VAD

VAD is our constructed **V**accine **A**ttitude **D**ataset. Following Hussain et al. [105], we crawl tweets using the Twitter streaming API with 60 pre-defined keywords relating to COVID-19 vaccines (e.g., *Pfizer*, *AstraZeneca*, and *Moderna*). We collected tweets between February 7th and April 3rd, 2022 using 60 vaccine-related keywords. The exhaustive list is: *'covid-19 vax'*, *'covid-19 vaccine'*, *'covid-19 vaccines'*, *'covid-19 vaccination'*, *'covid-19 vaccinations'*, *'covid-19 jab'*, *'covid-19 jabs'*, *'covid19 vax'*, *'covid19 vaccine'*, *'covid19 vaccines'*, *'covid19 vaccination'*, *'covid19 vaccinations'*, *'covid19 jab'*, *'covid19 jabs'*, *'covid vax'*, *'covid vaccine'*, *'covid vaccines'*, *'covid vaccination'*, *'covid vaccinations'*, *'covid jab'*, *'covid jabs'*, *'coronavirus vax'*, *'coronavirus vaccine'*, *'coronavirus vaccines'*, *'coronavirus vaccination'*, *'coronavirus vaccinations'*, *'coronavirus jab'*, *'coronavirus jabs'*, *'Pfizer vaccine'*, *'BioNTech vaccine'*, *'Oxford vaccine'*, *'AstraZeneca vaccine'*, *'Moderna vaccine'*, *'Sputnik vaccine'*, *'Sinovac vaccine'*, *'Sinopharm vaccine'*, *'Pfizer jab'*, *'BioNTech jab'*, *'Oxford jab'*, *'AstraZeneca jab'*, *'Moderna jab'*, *'Sputnik jab'*, *'Sinovac jab'*, *'Sinopharm jab'*, *'Pfizer vax'*, *'BioNTech vax'*, *'Oxford vax'*, *'AstraZeneca vax'*, *'Moderna vax'*, *'Sputnik vax'*, *'Sinovac vax'*, *'Sinopharm vax'*, *'Pfizer vaccinate'*, *'BioNTech vaccinate'*, *'Oxford vaccinate'*, *'AstraZeneca vaccinate'*, *'Moderna vaccinate'*, *'Sputnik* *vaccinate'*, *'Sinovac vaccinate'*, *'Sinopharm vaccinate'*.

Only English tweets were collected. Retweets were discarded. For preprocessing, hyperlinks, usernames and irregular symbols were removed. Emojis and emoticons were converted to their literal meanings using an emoticon dictionary[3]. The final dataset comprises 1.9 million English tweets. We randomly sample a subset of tweets for annotation. Upon an initial inspection, we found that over 97% of tweets mentioned only one aspect. As such, we annotate each tweet with a stance label and a text span characterizing the aspect. The annotation guideline comprises four questions:

- What is the stance towards vaccination?
- What is the Aspect Span? (i.e., Events or targets, it can be nouns, noun phrase, clause or sentence with verbal predicates).
- What is the opinion term/span? It should be opinion expressions, comprising both explicit and implicit expressions of stance.
- What is the Aspect category? It should be one of the pre-defined aspect categories (shown in Table 6.4).

The annotators have the choice to skip some of the questions if they find it difficult to answer. Taking the tweet '*Very grateful to those at Oxford. I've got my first #Covid19 vaccine.*' as an example, the annotators are expected to answer with: '*Pro-vaccine*', '*I've got my first #Covid19 vaccine*', '*Very grateful to those at Oxford. I've got my first #Covid19 vaccine*', '*2*'. If an annotator chooses to skip a tweet at any step of the process, this tweet will be recorded as skipped and the annotator will not be assigned with similar tweets. We first had a trial run where each annotator was asked to annotate the same set of tweets. Any disagreement was recorded and discussed to refine our annotation guideline in order to achieve consistency between the annotators.

In total, 2,800 tweets have been annotated in which 2,000 are used for training and the remaining 800 are used for testing. The statistics of the dataset is listed in Table 6.1. The stance labels are imbalanced. On the other hand, the average opinion length is longer than the average aspect length, and is close to the average tweet length. For the purpose of evaluation on tweet clustering and latent topic disentanglement, we further annotate tweets with a categorical label indicating the aspect category. Inspired by [189], we identify 24 aspect categories and each tweet is annotated with one of these categories. It is worth mentioning that aspect category labels are not used for training.

[3] `https://wprock.fr/en/t/kaomoji/`

	VAD		VC	
Specification	Train	Test	Train	Test
# tweets	2000	800	1162	531
# anti-vac.	638	240	822	394
# neutral	142	76	41	27
# pro-vac.	1220	484	299	110
Avg. length	33.5	34.13	29.6	30.24
len(aspect)	17.5	18.75	1.03	1.08
len(opinion)	27.97	29.01	3.25	3.15
# tokens	67k	27.3k	34.4k	16.8k

Table 6.1: Dataset Statistics. '# tweets' denotes the number of tweets in VAD, and for VC it is the number of sentences. 'anti-vac.' means *anti-vaccination* while 'pro-vac.' means *pro-vaccination*. 'Avg. length' and '# token' measure the number of word tokens.

VC

Vaccination **C**orpus [189] consists of 294 Internet documents about online vaccine debate annotated with events, 210 of which are annotated with opinions (in the form of text spans) towards vaccines. The stance label is considered to be the stance for the whole sentence. Those sentences with conflicting stance labels are regarded as neutral. We split the dataset into a ratio of 2:1 for training and testing. This eventually left us with 1,162 sentences for training and 531 sentences for testing.

6.4.2 Baselines

We compare the experimental results with the following baselines:
BertQA [146]: a pre-trained language model well-suited for span detection. With BertQA, attitude detection is performed by first classifying stance labels then predicting the answer queried by the stance label. The text span is configured as the ground-truth answer. We rely on its HuggingFace[4] [294] implementation. We employ ALBERT [132] as the backbone language model for both BertQA and VADet.
ASTE [208]: a pipeline approach consisting of aspect extraction [146] and sentiment labelling [145].

[4] `https://huggingface.co/transformers/model_doc/albert.html#albertforquestionanswering`

6.4.3 Hyper-parameters and Training Details

The dimensions of z_a, z_w and z_s are 768, 768 and 32, respectively. For each tweet, the number of samples from $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is 1. We modified the LM-fine-tuning script[5] from the HuggingFace library to implement VADET in the masked LM learning. We use default settings for the training script (i.e., Trainer in the HuggingFace library[6]), except for the batch size which is set to 128. The data pre-processor for the masked language model is the data collator for language modeling[7], which provides the function of randomly masking the tokens. The tokenizer for the data collator is the ready-to-use ALBERT tokenizer[8]. For the pre-trained language model (i.e., ALBERT) employed in this model, we inherit the default setting from the `AlbertConfig` class. We train VADET for 5 epochs on the un-annotated corpus.

In the supervised training of VADET, we use a batch size of 64. The learning rate is initialized to $2e-5$ with a linear warm-up schedule. We employ 5-fold training in which the training set is split into 5 subsets, of which 4 are used for training and the rest is for validation at the end of each epoch, and the final prediction is an ensemble of 5 independently-saved models. We train each model for 5 epochs, which takes roughly 2 hours on a node of a single Nvidia RTX 2080 GPU.

6.4.4 Evaluation Metrics

For stance classification, we use accuracy and Macro-averaged F1 score. For aspect span detection, we follow Rajpurkar et al. [223] in adopting exact match (EM) accuracy of the starting-ending position and Macro-averaged F1 score of the overlap between the prediction and ground truth aspect span. For tweet clustering, we follow Xie et al. [299] and Zhang et al. [313] and use the Normalized Mutual Information (NMI) metric to measure how the clustered group aligns with ground-truth categories. In addition, we also report the clustering accuracy.

In all our experiments, VADET is firstly pre-trained in an unsupervised way on our collected 1.9 million tweets before fine-tuning on the annotated training set from the VAD or VC corpora.

[5] `https://github.com/huggingface/transformers/blob/master/examples/pytorch/language-modeling/run_mlm.py`

[6] `https://huggingface.co/docs/transformers/master/en/main_classes/trainer#transformers.Trainer`

[7] `https://huggingface.co/docs/transformers/main_classes/data_collator`

[8] `https://huggingface.co/docs/transformers/master/en/model_doc/albert#transformers.AlbertTokenizer`

Model	VAD		VC	
Stance	Acc.	F1	Acc.	F1
BertQA	0.754	0.742	0.719	0.708
ASTE	0.723	0.710	0.704	0.686
VADET	**0.763**	**0.756**	**0.727**	**0.713**
Aspect Span	Acc.	F1	Acc.	F1
BertQA	0.546	0.722	0.525	0.670
ASTE	0.508	0.684	0.467	0.652
VADET	**0.556**	**0.745**	**0.541**	**0.697**
Cluster	Acc.	NMI	Acc.	NMI
DEC (BertQA)	0.633	58.1	0.586	52.8
K-means (BERT)	0.618	56.4	0.571	50.1
DEC (VADET)	**0.679**	**60.7**	**0.605**	**54.7**

Table 6.2: Results for stance classification, aspect span extraction and aspect clustering on both VAD and VC corpora.

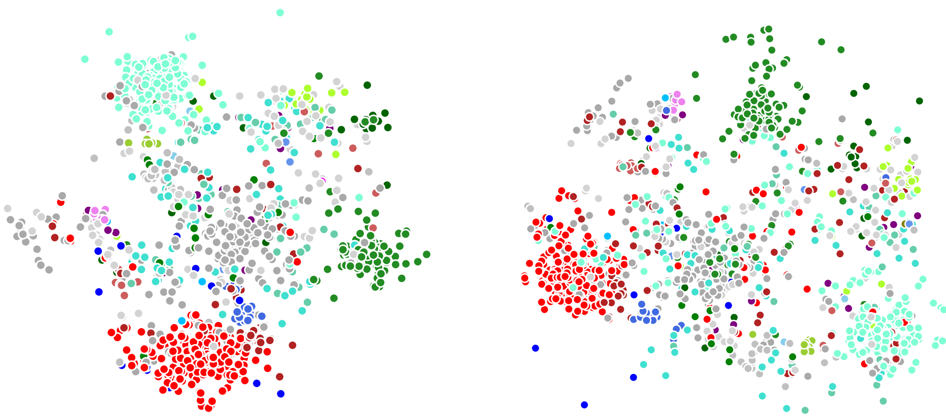
6.5 Experimental Results

6.5.1 Classification and Aspect Span Detection

In Table 6.2, we report the performance on attitude detection. In stance classification, our model outperforms both baselines with more significant improvements on ASTE. On aspect span extraction, VADET yields even more noticeable improvements, with a 2.3% increase in F1 over BertQA on VAD, and 2.7% on VC. These results indicate that the successful prediction relies on the hidden representation learned in the unsupervised training. The disentanglement of stance and aspect may have also contributed to the improvement.

6.5.2 Cluster Semantic Coherence Evaluation

To assess whether the learned latent aspect topics would allow meaningful categorization of documents into attitude clusters, we perform clustering using the disentangled representations that encode aspects, i.e., z_w. Deep Embedding Clustering (DEC) [299] is employed as the backend. For comparison, we also run DEC on the aspect representations of documents returned by BertQA. For each document, its aspect representation is obtained by averageing over the fine-tuned ALBERT representations of the constituent words in its aspect span. To assess the quality of clusters, we need the annotated aspect categories for documents in the test set. In VAD, we use the annotated aspect labels as the ground-truth categories whereas in VC we use the annotated event types. Results are presented in the lower part of Table 6.2. We found a prominent increase in NMI score over the baselines. Using the learned latent aspect topics as features, DEC (VADET) outperforms DEC (BertQA) by 4.6% and 1.9% in accuracy on VAD and VC, respectively. We also notice that using K-means as the clustering approach directly on the BERT-encoded tweet representations gives worse results compared to DEC. A similar trend is observed on the NMI metric. The improvements are shown visually in Figure 6.4 where the clustered groups produced by VADET are more identifiable. In the absence of categorical labels, the perspective expressed by each group can be inferred from the constituent tweets. For example, the tweet *'@user Georgian nurse dies of allergic reaction after receiving AstraZeneca Covid19 vaccine'* lies in the centroid of the red group, which relates to safety concerns.



(a) VADET

(b) BertQA

Figure 6.4: Clustered groups of VADET and BertQA on the VAD dataset. Each color indicates a ground truth aspect category. The clusters are dominated by: *(1) Red: the (adverse) side effects of vaccines; (2) Green: explaining personal experiences with any aspect of vaccines; and (3) Cyan: the immunity level provided by vaccines.*

We also evaluate the semantic coherence of the clustered tweets. The semantic coherence is the extent to which tweets within a cluster belong to each other, which is employed as an evaluation metric for cluster quality evaluation in an unsupervised way. Recent work of Bilal et al. [21] found that Text Generation Metrics (TGMs) align well with human judgement in evaluating clusters in the context of microblog posts. TGM by definition measures the similarity between the ground-truth and the generated text. The rationale is that a high TGM score means sentence pairs are semantically similar. Here, two metrics are used: *BERTScore*, which calculates the similarity of two sentences as a sum of cosine similarities between their tokens' embeddings [318], and *BLEURT*, a pre-trained adjudicator that fine-tunes BERT on an external dataset of human ratings [239]. As in [21], we adopt the Exhaustive Approach that for a cluster C, its coherence score is the average TGM score of every possible tweet pair in the cluster:

$$f(C) = \frac{1}{N^2} \sum_{i,j \in [1,N], i<j} \mathrm{TGM}(\mathrm{tweet}_i, \mathrm{tweet}_j).$$

Figure 6.5 shows the BERTScore and the BLEURT score of VADET and baselines on two datasets. The VADET shows consistent improvements across the datasets. This indicates that tweets clustered using the latent aspect topics generated by VADET are semantically more similar, thus validating the assumption that disentangled representations are more effective in bringing together tweets of a similar gist.

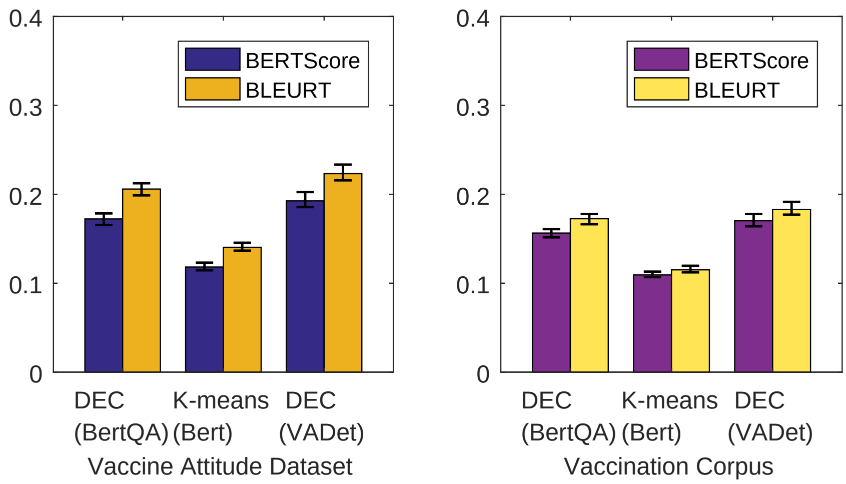


Figure 6.5: Semantic coherence evaluated in two metrics.

6.5.3 Ablations

We conduct ablation studies to investigate the effect of semi-supervised learning that uses the variational latent representation learning approach and aspect-stance disentanglement on the latent semantics. We study their effects on stance classification and aspect span detection. The results are reported in Table 6.3.

We can observe that on VAD without disentangled learning or unsupervised pre-training results in the degradation of the stance classification performance. How-

Model	VAD		VC	
Stance	Acc.	F1	Acc.	F1
VADET	**0.763**	**0.756**	0.727	0.713
VADET-D	0.751	0.746	**0.736**	**0.716**
VADET-U	0.741	0.734	0.712	0.698
Aspect Span	Acc.	F1	Acc.	F1
VADET	**0.556**	**0.745**	**0.541**	**0.697**
VADET-D	0.540	0.728	0.537	0.684
VADET-U	0.528	0.712	0.525	0.653

Table 6.3: Results of stance classification and aspect span detection of VADET without disentanglement (-D) or unsupervised pre-training (-U).

ever, on VC, we see a slight increase in classification accuracy without disentangled learning. We attribute this to the vagueness of the stance which might cause the model to disentangle more than it should be. On the aspect span detection task, we observe consistent performance drop across all metrics and on both datasets. In particular, without the pre-training module, the performance drops more significantly. These results indicate that semi-supervised learning is highly effective with VAE, and the disentanglement of stance and aspect serves as a useful component, which leads to noticeable improvements.

6.6 Summary

This chapter presents a semi-supervised model to detect user attitudes and distinguish aspects of interest in vaccines on social media. We employed a Variational Auto-Encoder to encode the main topical information into the language model by unsupervised training on a massive, unannotated dataset. The model is then further trained under a semi-supervised setting that leverages annotated stance labels and aspect spans to induce the disentanglement between stances and aspects in a latent semantic space. We empirically showed the benefits of such an approach for attitude detection and aspect clustering over two vaccine corpora. Ablation studies show that disentangled learning and unsupervised pre-training are important to effective vaccine attitude detection. Further investigations on the quality of the disentangled representations verify the effectiveness of the disentangled factors.

Label	Definition
1	AstraZeneca: How health organisations/institution, communities, groups, individuals and other entities position themselves towards vaccines
2	AstraZeneca: Explaining personal experiences with any aspect of vaccines
3	AstraZeneca: The achievement that vaccines have brought (vaccines save lives, protect the community, protect future generations)
4	AstraZeneca: The (adverse) side effects of vaccines: illnesses, symptoms, deaths
5	AstraZeneca: The immunity level provided by vaccines
6	AstraZeneca: The economic effect of vaccination (less illnesses, less expenses for family and society)
7	AstraZeneca: Discussing the personal freedom to choose in relation to vaccines
8	AstraZeneca: Discussing the relation between vaccines and religion, conspiracy or moral attitudes
9	Pfizer or Moderna: How health organisations/institution, communities, groups, individuals and other entities position themselves towards vaccines
10	Pfizer or Moderna: Explaining personal experiences with any aspect of vaccines
11	Pfizer or Moderna: The achievement that vaccines have brought (vaccines save lives, protect the community, protect future generations)
12	Pfizer or Moderna: The (adverse) side effects of vaccines: illnesses, symptoms, deaths
13	Pfizer or Moderna: The immunity level provided by vaccines
14	Pfizer or Moderna: The economic effect of vaccination (less illnesses, less expenses for family and society)
15	Pfizer or Moderna: Discussing the personal freedom to choose in relation to vaccines
16	Pfizer or Moderna: Discussing the relation between vaccines and religion, conspiracy or moral attitudes

17	Other Brands or not mentioned: How health organisations/institution, communities, groups, individuals and other entities position themselves towards vaccines
18	Other Brands or not mentioned: Explaining personal experiences with any aspect of vaccines
19	Other Brands or not mentioned: The achievement that vaccines have brought (vaccines save lives, protect the community, protect future generations)
20	Other Brands or not mentioned: The (adverse) side effects of vaccines: illnesses, symptoms, deaths
21	Other Brands or not mentioned: The immunity level provided by vaccines
22	Other Brands or not mentioned: The economic effect of vaccination (less illnesses, less expenses for family and society)
23	Other Brands or not mentioned: Discussing the personal freedom to choose in relation to vaccines
24	Other Brands or not mentioned: Discussing the relation between vaccines and religion, conspiracy or moral attitudes

Table 6.4: The predefined aspect categories and their definitions.

Chapter 7

Disentangling Aspect and Stance via a Siamese Autoencoder for Aspect Clustering

Chapter Abstract

In this chapter, we build models to disentangle the aspect and the stance in the task of vaccination opinion mining. The disentangled representation enables us to cluster tweets based on aspect similarity rather than sentence similarity, allowing the model to deal with unseen tweets more effectively. We first use a denoising autoencoder built on a pre-trained language model to capture the vaccine-related topics from myriads of unlabelled tweets. We then enable the disentanglement of the latent space by using biases from stance labels and aspect text spans handled by the disentangled cross attention. Finally, we introduce the Swapping Autoencoder to align evidence of stance and aspect to latent vectors by swapping the presumed aspect embedding of a tweet with that of another discussing the same aspect. The three components are integrated into a clustering-friendly representation learning method that produces disentangled representations for aspect-oriented clustering of tweets. In experiments on two Twitter vaccination corpora, we show that the model discovered disentangled representations which improved clustering results. With a classification head for inductive biases, the model can make stance predictions comparable to backbone language

models.

7.1 Introduction

Mining public opinions about vaccines from social media has been hindered by the wide variety of users' attitudes, and the continuously new aspects arising in the public debate of vaccination [106]. The most recent approaches have adopted holistic frameworks built on morality analysis [200] or neural-based models predicting users' stances on different aspects of the online debate [334]. So far, these frameworks have been frequently framed via well-known tasks, such as aspect classification or text span detection, that use supervision to train text classifiers. However, such a direct usage of the supervision information has constrained the models to predefined aspect classes and restricted their flexibility in generalising to opinions with aspects never seen before (e.g., new moral issues or immunity level).

To mitigate this limitation, some of the most promising approaches have been devised as supervised models generating *clustering-friendly representations* [266]. These have recently shown promising results on open-domain tasks when combined with pre-trained language models (PLM) thanks to their flexibility, generalisation, and need for minimal tweaks [225, 252]. However, despite the improved capabilities in capturing the overall text semantics, existing models for text clustering [177, 184, 246, 313] still struggle to distinguish between the mixed users' stances and aspects on vaccination, and as a result, they often generate clusters that do not reflect the novel aspects of interest. As an illustrating example, consider the tweets *"mRNA vaccines are poison"* and *"The Pfizer vaccine is safe"*, that the majority of existing methodologies are prone to cluster into different groups due to the opposite stances manifested, despite the fact that both of them are targeting safety issues.

To address the aforementioned problem, we posit that a model should be able to (i) disentangle the stance from the aspect discussed, and simultaneously (ii) use the generated representations in a framework (e.g., clustering) that ease the integration of aspects never seen before. We thus propose a novel representation learning approach, called the *Disentangled Opinion Clustering* (DOC) model, which performs disentangled learning [172] via text autoencoders [31, 187], and generates *cluster-friendly* representations suitable for the integration of novel aspects. The proposed model, DOC, learns clustering-friendly representations through a denoising autoencoder [187] driven by out-of-the-box Sentence-BERT embeddings [225], and disentangles stance from opinions by using the supervision signal to drive a disentangled cross-attention mechanism and a Swapping Autoencoder[204].

We conducted an experimental assessment on two publicly available datasets on vaccination opinion mining, the Covid-Moral-Foundation (CMF) dataset [200] and the Vaccination Attitude Detection (VAD) corpora [334]. We first assessed the quality of the disentangled representation in generating aspect-coherent clusters. Then, we measured the generalisation of the proposed approach via a cross-dataset evaluation by performing clustering on a novel dataset with unknown aspect categories. Finally, we showed the benefit of this approach on the traditional stance classification task, along with a report on the thorough ablation study highlighting the impact of each model component on the clustering quality and the degree of disentanglement of the generated representations.

7.2 Related Work

The proposed work is related to sentence bottleneck representations, disentangled latent representations, clustering in NLP and vaccination opinion mining.

Sentence Bottleneck Representation Sentence representation learning typically aims to generate a fixed-sized latent vector that encodes a sentence into a low-dimensional space. In recent years, in the wake of the wide application of pre-trained language models (PLMs), several approaches have been developed leveraging the pre-trained information to encode sentence semantics. The most prevalent work is the SBERT [225] that fine-tunes BERT [60] on the SNLI dataset [30] through a siamese pooling structure. The learned representations are immediately applicable to a wide range of tasks, such as information retrieval and clustering, significantly reducing the effort required to generate the task-specific representations [268]. More recently, Montero et al. [187] presented a sentence bottleneck autoencoder, called AutoBot, that learns a latent code by reconstructing the perturbated text. Their model indicates the importance of topic labels as reconstruction objectives.

Disentangled Latent Representation Earlier works explored disentangled representation to facilitate domain adaptation [19, 125, 172]. In recent years, John et al. [115] generated disentangled representations geared to transfer holistic style such as tone and theme in text generation. Park et al. [204] proposed the Swapping autoencoder to separate texture encoding from structure vectors in image editing. The input images are formed in pairs to induce the model to discern the variation (e.g., structure) while retaining the common property (e.g., texture). However, recent studies show that disentanglement in the latent space is theoretically unachievable without access to some inductive bias [157]. It is suggested that local isometry between variables of interest is sufficient to establish a connection between the observed variable and the latent variable [97, 158], even with few annotations [159]. This is in line with [225] which illuminates our work to utilize labels and reconstruction of perturbed text to induce the disentanglement.

Text Clustering The recent development in neural architectures has reshaped clustering practices [299]. For example, Zhang et al. [316] leveraged transformer encoders for clustering over the user intents. Several methods utilised PLM embeddings to discover topics which were subsequently used for clustering news articles and product reviews [104, 178]. Others exploited the neural components, i.e., the BiLSTM-CNN [315], the CNN-Attention [79] and the Self-Attention [319] to offer end-to-end clustering. Zhang et al. [313] developed the Supporting Clustering with Contrastive Learning (SCCL) model by augmenting the disparity between short text. A notable work is DS-Clustering [252], which extracts aspect phrases first and then clusters the aspect embeddings. Outside of clustering methods, there is a surging interest in clustering-friendly representation learning [266]. Yet, few methods cluster documents along a particular axis or provide disentangled representations to cluster over a subspace.

Vaccination Opinion Mining The task of vaccination opinion mining is commonly carried out on social media to detect user attitudes and provide insights to be used against the related 'infodemic' [38, 130, 289, 320]. Recent approaches rely on semantic matching and stance classification with extensions including human-in-the-loop protocols and text span prediction to scale to the growing amount of text [200, 334].

7.3 Disentangled Opinion Clustering Model

We build our approach upon two vaccination opinion corpora [200, 334]. In both corpora, a small number of tweets are labelled, each of which is annotated with a stance label ('*pro-vaccine*', '*anti-vaccine*' and '*neutral*') and a text span or an argumentative pattern denoting an aspect. For example, for the tweet, '*The Pfizer vaccine is safe.*', its stance label is '*pro-vaccine*' and the argumentative pattern is '*vaccine safety*'. Since vaccination opinions explode over time, supervised classifiers

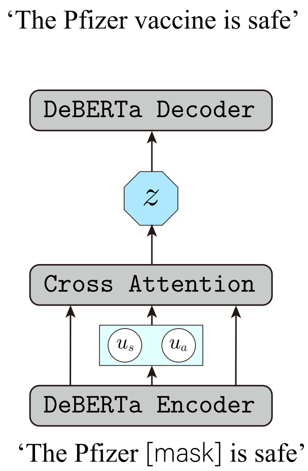
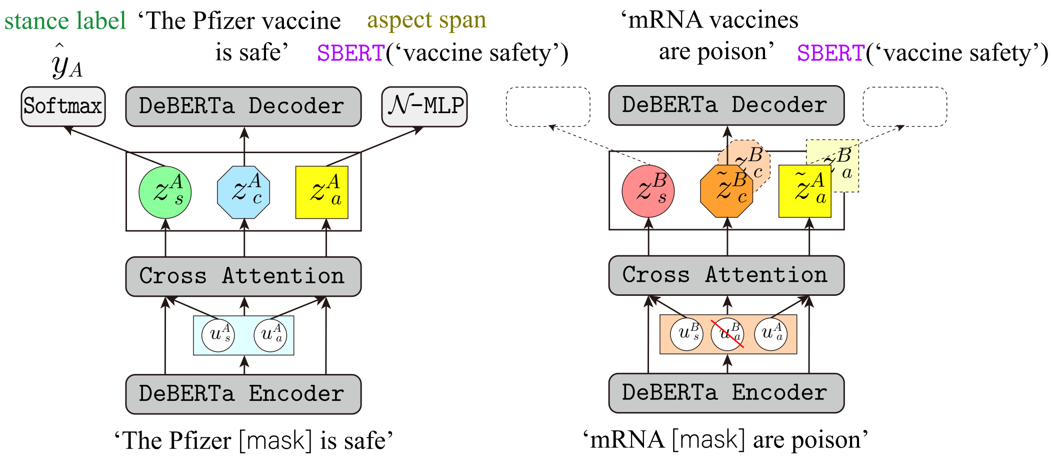


Figure 7.1: Disentangled Opinion Clustering (DOC) model in unsupervised learning. A tweet is fed into an autoencoder with DeBERTa as both the encoder and decoder to learn the latent sentence vector $\boldsymbol{z}$.

or aspect extractors would soon become outdated and fail to handle constantly evolving tweets. In an effort to mitigate this issue, we address the problem of vaccination opinion mining by learning disentangled stance and aspect vectors of tweets in order to cluster tweets along the aspect axis.

Our proposed model, called Disentangled Opinion Clustering (DOC), is shown in Figure 7.1 - 7.2. It is trained in two steps. In **unsupervised learning** (Figure 7.1), a tweet is fed into an autoencoder with DeBERTa as both the encoder and the decoder to learn the latent sentence vector $\boldsymbol{z}$. Here, each tweet is mapped to two embeddings, the context embedding $\boldsymbol{u}_s$ which encodes the stance label information and the aspect embedding $\boldsymbol{u}_a$ which captures the aspect information. Under unsupervised learning, these two embeddings are not distinguished. Together with the hidden representation of the input text, H, they are mapped to the latent sentence vector $\boldsymbol{z}$ by cross-attention. As the autoencoder can be trained on large-scale unannotated tweets relating to vaccination, it is expected that $\boldsymbol{z}$ would capture the vaccine-related topics.

Then in the second step of **supervised learning** (Figure 7.2), the DeBERTa-based autoencoder is fine-tuned to learn the latent stance vector $\boldsymbol{z}_s$ and the latent aspect vector $\boldsymbol{z}_a$ using the tweet-level annotated stance label and aspect text span (or the argumentative pattern ‘*vaccine safety*’ in Figure 7.2) as the inductive bias. Here, the latent stance vector $\boldsymbol{z}_s$ is derived from $\boldsymbol{u}_s$. It is expected that $\boldsymbol{z}_s$ can be used to predict the stance label. On the other hand, the latent aspect vector $\boldsymbol{z}_a$ is derived from $\boldsymbol{u}_a$ only, and it can be used to generate the SBERT-encoded aspect



(a) Disentanglement with Inductive Biases (b) Swapped Auto-Encoder

Figure 7.2: Disentangled Opinion Clustering (DOC) Model in supervised learning. **(a) Disentanglement with inductive biases.** The DeBERTa-based autoencoder is fine-tuned to learn the latent stance vector $\boldsymbol{z}_s$ and the latent aspect vector $\boldsymbol{z}_a$ using the tweet-level annotated stance label and aspect text span (or the argumentative pattern '*vaccine safety*' for the input tweet) as the inductive bias; **(b) Swapping autoencoder.** To enable a better disentanglement of $\boldsymbol{z}_s$ and $\boldsymbol{z}_a$, for the two tweets discussing the same aspect but with different stance labels, tweet B's aspect embedding $\boldsymbol{u}_a^B$ is replaced by the tweet A's aspect embedding $\boldsymbol{u}_a^A$. As the two tweets discuss the same aspect, their aspect embeddings are expected to be similar. As such, we can still reconstruct tweet B using the latent content vector $\boldsymbol{z}_c^B$ derived from the swapped aspect embedding. Note that (a) and (b) are learned simultaneously.

text span. Both $\boldsymbol{z}_s$ and $\boldsymbol{z}_a$, together with the hidden representation of the input text H, are used to reconstruct the original text through the DeBERTa decoder. The training instances are organized in pairs since we use the idea of swapped autoencoder (shown in Figure 7.2(b)) to swap the aspect embedding of one tweet with that of another if both discuss the same aspect. The resulting latent vector can still be used to reconstruct the original tweet. In what follows, we describe the two steps, unsupervised and supervised learning, in detail.

7.3.1 Unsupervised Learning of Sentence Representation

Due to the versatility of PLMs, sentence representations are usually derived directly from contextualised representations generated by the PLMs. However, as has been previously discussed in Montero et al. [187], sentence representations derived in this way cannot guarantee reliable reconstruction of the input text and are therefore less suitable for efficient conditional text generation. Partly inspired by the use of autoencoder for sentence representation learning as in [187], we adopt the autoencoder architecture to initially guide the sentence representation learning by fine-tuning it on vaccination tweets. Rather than RoBERTa [155], we adopt DeBERTa, a variant of BERT in which each word is represented using two vectors encoding its content and position. The attention weight of a word pair is computed as a sum of four attention scores calculated from different directions based on their content/position vectors, i.e., content-to-content, content-to-position, position-to-content, and position-to-position. Instead of representing each word by a content vector and a position vector, we modify DeBERTa by representing an input sentence using two vectors, a context embedding $\boldsymbol{u}_s$ encoding its stance label information and an aspect embedding $\boldsymbol{u}_a$ encoding its aspect information. We will discuss later in this section how to perform disentangled representation learning with $\boldsymbol{u}_s$ and $\boldsymbol{u}_a$. During the unsupervised learning stage, we do not distinguish between $\boldsymbol{u}_s$ and $\boldsymbol{u}_a$ and simply use $\boldsymbol{u} = [\boldsymbol{u}_s, \boldsymbol{u}_a]$ to denote them.

More specifically, we train the autoencoder AUTOBOT on an unannotated Twitter corpus with the masked token prediction as the training objective. The encoder applies the multi-head attention to clamp the hidden representations of the top layer of the pre-trained transformer. If we use H to denote the hidden representations, the multi-head attention can be expressed as:

$$\text{head}_i = \text{softmax}\left(\frac{\mathbf{u}W_Q(HW_K)^\top}{\sqrt{d_H}}\right)HW_V, \tag{7.3.1}$$

$$\mathbf{z} = [\text{head}_1, \text{head}_2, \ldots, \text{head}_h]W_O, \tag{7.3.2}$$

where $H \in \mathbb{R}^{n \times d_H}$, $W_Q \in \mathbb{R}^{2d_H \times d_K}, W_K \in \mathbb{R}^{d_H \times d_K}$, $W_V \in \mathbb{R}^{d_H \times d_V}$, $\text{head}_i \in \mathbb{R}^{d_V}$ and $W_O \in \mathbb{R}^{hd_V \times d_{\mathbf{z}}}$. $\mathbf{u} \in \mathbb{R}^{2d_H}$ is generated from a fully-connected layer over the hidden vectors. The bottleneck representation $\mathbf{z}$ is supposed to encode the semantics of the whole sentence.

The transformer decoder comprises n layers of cross-attention such that the output of the previous layer is processed by a gating mechanism [96]. The recurrence is repeated n times to reconstruct the input, where n denotes the token length of the input text.

7.3.2 Injecting Inductive Biases by Disentangled Attention

Recent work on disentanglement learning suggested unsupervised disentanglement is impossible without inductive bias [159]. In the datasets used in our experiments, there are a small number of labelled tweets. We can use the tweet-level stance labels and the annotated aspect text spans as inductive bias. Here, the disentangled attention of DeBERTa is utilized to mingle different factors. Assuming each sentence is mapped to two vectors, the context vector $\boldsymbol{u}_s$ encoding its stance label information and the aspect vector $\boldsymbol{u}_a$ encoding its aspect information, we can then map $\boldsymbol{u}_s$ to a latent stance vector $\boldsymbol{z}_s$ which can be used to predict the stance label, and map $\boldsymbol{u}_a$ to a latent aspect vector $\boldsymbol{z}_a$ which can be used to reconstruct the aspect text span. We use the cross-attention between $\boldsymbol{u}_s$ and $\boldsymbol{u}_a$ to reconstruct the original input sentence.

Stance Classification Let $\mathbf{h}_{\texttt{CLS}}$ denote the hidden representation of the `[CLS]` token, the stance bias is injected by classification over the stance categories:

$$\mathbf{z}_{\boldsymbol{s}} = \text{softmax}\left(\frac{\mathbf{u}_{\boldsymbol{s}} W_{\boldsymbol{q},\boldsymbol{s}} (\mathbf{h}_{\texttt{CLS}} W_{\boldsymbol{k},\texttt{CLS}})^{\top}}{\sqrt{d_H}}\right) \mathbf{h}_{\texttt{CLS}} W_{\boldsymbol{v},\texttt{CLS}}, \tag{7.3.3}$$

$$\hat{y}_s = \text{softmax}(\mathbf{z}_{\boldsymbol{s}} W), \quad \mathcal{L}_s = -y_s^{(i)} \log \hat{y}_s^{(i)}. \tag{7.3.4}$$

Essentially, we use $\boldsymbol{u}_s$ as query and $\mathbf{h}_{\texttt{CLS}}$ as key and value to derive $\boldsymbol{z}_s$, which is subsequently fed to a softmax layer to predict a stance label $\hat{y}_s$. The objective function for stance classification is a cross-entropy loss between the true and the predicted labels.

Aspect Text Span Reconstruction We assume $\boldsymbol{u}_a$ encoding the sentence-level aspect information and use self-attention to derive the latent aspect representation $\boldsymbol{z}_a$. To reconstruct the aspect text span from $\boldsymbol{z}_a$, we use the embedding generated by SBERT [225] as the targeted aspect text span embedding since SBERT has been empirically shown achieving the state-of-the-art on Semantic Textual Similarity (STS) tasks. Those clustering-friendly representations, if they encode the argumentative patterns or aspect spans alone, are strong inductive biases in the axis of aspects.

Specifically, the sentence embedding of the aspect expression is generated by a Gaussian MLP decoder [123]:

$$\mathbf{z}_a = \text{softmax}\left(\frac{\mathbf{u}_a W_{\boldsymbol{q},\boldsymbol{a}} (\mathbf{u}_a W_{\boldsymbol{k},\boldsymbol{a}})^{\top}}{\sqrt{d_H}}\right) \mathbf{u}_a W_{\boldsymbol{v},\boldsymbol{a}}, \tag{7.3.5}$$

$$\mathcal{L}_a = -\log \mathcal{N}(\mathbf{y}_a; \text{MLP}_{\mu}(\mathbf{z}_{\boldsymbol{a}}), \text{MLP}_{\sigma}(\mathbf{z}_{\boldsymbol{a}})\mathbf{I}), \tag{7.3.6}$$

where $\boldsymbol{x}_a$ denotes the aspect text span in the original input sentence, $\mathbf{y}_a$ is the ground-truth aspect text span embedding produced by $\mathbf{y}_a = \texttt{SBERT}(\boldsymbol{x}_a)$, whose value is used for computing the Gaussian negative log-likelihood loss[1].

Input Text Reconstruction To reconstruct the original input text, we need to make use of both the latent stance vector $\boldsymbol{z}_s$ and the latent aspect vector $\boldsymbol{z}_a$. Here we use the cross attention of these two vectors to derive the content vector $\boldsymbol{z}_c$.

$$\begin{aligned}
Q_{\boldsymbol{c}} = \mathbf{u}W_{\boldsymbol{q},\boldsymbol{c}}, \quad K_{\boldsymbol{c}} = HW_{\boldsymbol{k},\boldsymbol{c}}, \quad V_{\boldsymbol{c}} = HW_{\boldsymbol{v},\boldsymbol{c}}, \\
Q_{\boldsymbol{s}} = \mathbf{u}_{\boldsymbol{s}}W_{\boldsymbol{q},\boldsymbol{s}}, \quad K_{\boldsymbol{s}} = \mathbf{u}_{\boldsymbol{s}}W_{\boldsymbol{k},\boldsymbol{s}}, \\
Q_{\boldsymbol{a}} = \mathbf{u}_{\boldsymbol{a}}W_{\boldsymbol{q},\boldsymbol{a}}, \quad K_{\boldsymbol{a}} = \mathbf{u}_{\boldsymbol{a}}W_{\boldsymbol{k},\boldsymbol{a}}, \\
a_j = Q_{\boldsymbol{c}}K_j^{\boldsymbol{c}\top} + Q_{\boldsymbol{c}}K_{\boldsymbol{s}}^{\top} + K_j^{\boldsymbol{c}}Q_{\boldsymbol{s}} + Q_{\boldsymbol{c}}K_{\boldsymbol{a}}^{\top} + K_j^{\boldsymbol{c}}Q_{\boldsymbol{a}} \\
\text{head}_i = \text{softmax}\left(\frac{\boldsymbol{a}}{\sqrt{5d_H}}\right)HW_{\boldsymbol{v},\boldsymbol{c}}, \\
\mathbf{z}_{\boldsymbol{c}} = [\text{head}_1, \text{head}_2, \ldots, \text{head}_h]W_O,
\end{aligned} \tag{7.3.7}$$

where $\mathbf{u} = [\mathbf{u}_s, \mathbf{u}_a]$, a_j is the j-th element of $\boldsymbol{a}$, and $K_j^{\boldsymbol{c}}$ represents the j-th row of $K_{\boldsymbol{c}}$. The resulting $\mathbf{z}_c$ is the content representation for reconstructing the original sentence.

7.3.3 Disentanglement of Aspect and Stance

Although the inductive biases, i.e., the tweet-level stance labels and annotated aspect text spans, are used to learn the latent stance vectors $\boldsymbol{z}_s$ and the aspect vectors $\boldsymbol{z}_a$ as discussed in the last subsection, there could still be possible dependences between the two latent variables. To further the disentanglement, we propose to swap the learned aspect embeddings of two tweets discussing the same aspect in Siamese networks. We draw inspiration from the Swapping Auto-Encoder [204] where a constituent vector of a Generative Adversarial Network (GAN) is swapped with that produced by another image. The original swapping autoencoder was designed for image editing and required a patch discriminator with texture cropping to the corresponding disentangled factors with the desired properties. In our scenario, such alignment is instead induced by tweets discussing the same aspect.

We create pairs of tweets by permutations within the same aspect group $\{\boldsymbol{x}^A, \boldsymbol{x}^B\}_{A,B \in G_k, A \neq B}$. Here, by abuse of notation, we use k to denote the k-th aspect group, G_k. The groups are identified by tweets with the same aspect label, regardless of their stances. We sketch the structure of pair-wised training in Figure 7.2(b). The tweets are organized in pairs and a bottleneck representation is obtained for each

[1] https://pytorch.org/docs/stable/generated/torch.nn.GaussianNLLLoss.html

tweet:

$$\mathbf{z}^A = \text{enc}(\boldsymbol{x}^A), \quad \mathbf{z}^B = \text{enc}(\boldsymbol{x}^B). \tag{7.3.8}$$

We would like $\mathbf{z}^A$ to disentangle into latent factors, i.e., the variation in a factor of $\mathbf{z}^A$ is associated with a change in $\boldsymbol{x}^A$ [158]. Unlike the majority of work [115, 322] that directly splits $\mathbf{z}^A$ in the latent space, we assume that the entangled vector is decomposed by a causal network. We train a vector $\mathbf{u} = [\mathbf{u}_s, \mathbf{u}_a]$ to trigger the activation of the networks (i.e., the self-attentions in Eq. 7.3.3-Eq. 7.3.7). The outputs of the networks are independent components that encode the desiderata. If $\mathbf{z}_s$ and $\mathbf{z}_a$ are parameterized independent components triggered by $\mathbf{u}_s$ and $\mathbf{u}_a$ respectively, the substitution of $\mathbf{u}_a^B$ with $\mathbf{u}_a^A$ can be regarded as soft exchanges between $\mathbf{z}_a^A$ and $\mathbf{z}_a^B$.

Based on this provisos, we substitute $\mathbf{u}_a^B$ with $\mathbf{u}_a^A$ to cause changes in $\mathbf{z}_c^B$. This substitution will also be reflected by changes in $\mathbf{z}_a^B$. In practice, we train on all permutations with the same aspect group, regardless of the stance. The reconstruction loss for each latent factor (i.e., stance and aspect) is calculated once to balance the number of training examples unless it is content text generated from the swapped bottleneck representation.

Formally, the Swapping Auto-Encoder presented in Figure 7.2(b) can be expressed as

$$\begin{aligned}
Q_{\boldsymbol{s}}^B &= \mathbf{u}_{\boldsymbol{s}}^B W_{\boldsymbol{q},\boldsymbol{s}}, \quad K_{\boldsymbol{s}}^B = \mathbf{u}_{\boldsymbol{s}}^B W_{\boldsymbol{k},\boldsymbol{s}}, \\
Q_a^A &= \mathbf{u}_a^A W_{\boldsymbol{q},\boldsymbol{a}}, \quad K_a^A = \mathbf{u}_a^A W_{\boldsymbol{k},\boldsymbol{a}}, \\
a_j &= Q_{\boldsymbol{c}} K_j^{\boldsymbol{c}\top} + Q_{\boldsymbol{c}} K_{\boldsymbol{s}}^{B\top} + K_j^{\boldsymbol{c}} Q_s^B + Q_{\boldsymbol{c}} {K_a^A}^\top + K_j^{\boldsymbol{c}} Q_a^A, \\
\text{head}_i &= \text{softmax}\left(\frac{\boldsymbol{a}}{\sqrt{5d_H}}\right) H W_{\boldsymbol{v},\boldsymbol{c}}, \\
\mathbf{z}_c^B &= [\text{head}_1, \text{head}_2, \ldots, \text{head}_h] W_O,
\end{aligned}$$

$$\begin{aligned}
\mathbf{z}_s^B &= \text{softmax}\left(\frac{\mathbf{u}_s^B W_{\boldsymbol{q},\boldsymbol{s}} (K_{\texttt{CLS}})^\top}{\sqrt{d_H}}\right) V_{\texttt{CLS}}, \\
\mathbf{z}_a^B &= \text{softmax}\left(\frac{Q_a^A (K_a^A)^\top}{\sqrt{d_H}}\right) \mathbf{u}_a^A W_{\boldsymbol{v},\boldsymbol{a}},
\end{aligned}$$

where $\mathbf{z}_c^B$ is input to the decoder for the reconstruction of $\boldsymbol{x}^B$. Note that the above equations are specially used in the swapping autoencoder for the computation of $\mathbf{z}^B$. If there is no substitution in the latent space, the above equations will not be calculated.

Aspect Group	Pro-Vax	Anti-Vax	Neutral
	CMF		
Care/Harm	70	11	2
Fairness/Cheating	25	18	13
Loyalty/Betrayal	25	0	5
Authority/Subversion	20	46	13
Purity/Degradation	2	15	0
Liberty/Oppression	6	62	5
Non-moral	167	47	41
	VAD		
Health Institution	400	84	36
Personal Experience	381	16	3
Vaccines Save Lives	12	1	0
(Adverse) Side Effects	179	256	63
Immunity Level	433	113	52
Economic Effects	23	12	5
Personal Freedom	5	18	7
Moral Attitudes	5	43	2

Table 7.1: Dataset statistics of CMF and VAD. We list the number of pro-vaccine, anti-vaccine and neutral tweets in each group.

Given $\mathcal{L}_c^B = \text{dec}(\mathbf{z}_c^B)$, the final objective function is written as

$$\mathcal{L} = \mathcal{L}_c^A + \lambda_s \mathcal{L}_s^A + \lambda_a \mathcal{L}_a^A + \lambda_B \mathcal{L}_c^B, \tag{7.3.9}$$

where λ_s, λ_a and λ_B are hyper-parameters controlling the importance of each desirable property. In our experiments we choose $\lambda_s = \lambda_a = 1$ and $\lambda_B = 0.5$.

7.4 Experimental Setup

7.4.1 Datasets

Statistics We conduct our experimental evaluation on two publicly available Twitter datasets about the Covid-19 vaccination: the Covid Moral Foundation (CMF) dataset [200] and the Vaccination Attitude Detection (VAD) corpus [334]. CMF is a tweet dataset focused on the Covid-19 vaccine debates, where each tweet is assigned an argumentative pattern. VAD consists of 8 aspect categories further refined by vaccine bands. Similar to the argumentative pattern in the CMF dataset, each tweet is characterised by a text span indicating its aspect. The dataset statistics are reported in Table 7.1, with examples shown in 7.2. The train/test split follows

4 : 1. For the unsupervised pre-training of sentence bottleneck representations, we combine the unlabelled Covid-19 datasets from both CMF[2] and VAD[3] repositories. The final dataset consists of 4.37 million tweets.

Format In the Covid-Moral-Foundation (CMF) dataset, each tweet is associated with a pre-defined and manually annotated argumentative pattern. The annotated tweets are categorized by moral foundations that can be regarded as coarse aspects distilled from argumentative patterns. Each moral foundation is associated with two polarities (e.g., *care/harm*), and is treated as the group label of a cluster of tweets. The polarity is given by the vaccination stance label. Among the examples in Table 7.2, *'The vaccine is safe'* is the argumentative pattern, while *'Care/Harm'* is the categorical label denoting the aspect group. An exhaustive list of the argumentative patterns can be found in the original paper of Pacheco et al. [200].

In Vaccination Attitude Detection (VAD), a training instance comprises a stance label, a categorical aspect label and an aspect text span. For example, Table 7.2 shows the tweet *'Study reports Oxford/AstraZeneca vaccine is protective against Brazilian P1 strain of COVID19.'* is annotated with the text span *'Oxford/AstraZeneca vaccine is protective against Brazilian P1 strain of COVID19'*, and its aspect belongs to the aspect category *'Immunity Level'*.

7.4.2 Baselines

We employ 5 baseline approaches: SBERT[4], AutoBot[5], DS-Clustering, VADet[6], and SCCL[7], of which SBERT and AutoBot are sentence embedding-approaches capturing the sentence-level semantic distance or similarity. VADet also learns disentangled representations. However, it is noteworthy that it employed DEC [299] as the clustering algorithm, and here we test its representations on distance-based clustering. SCCL performs joint representation learning and document clustering. DS-Clustering is a pipeline approach that predicts a text span and employs SBERT to generate an aspect embedding. For clustering-friendly representation learning methods, we examine their performance using k-means and its variant k-medoids [136], and the Agglomerative Hierarchical Clustering (AHC). The comparison involves three tasks: tweet clustering based on aspect categories (intra- and cross-datasets),

[2]`https://gitlab.com/mlpacheco/covid-moral-foundations`
[3]`https://github.com/somethingx1202/VADet`
[4]`https://github.com/UKPLab/sentence-transformers`
[5]`https://github.com/ivanmontero/autobot`
[6]`https://github.com/somethingx1202/VADet`
[7]`https://github.com/amazon-research/sccl`

CMF		
Tweet	**Argumentative Pattern**	**Aspect Group**
Vaccine decreases your chances of getting severe life-threat.	The vaccine is safe	Care/Harm
There is no way someone can tell me that the COVID vaccine does not cause harm to pregnant women.	The covid vaccine is harmful for pregnant women and kids	Care/Harm
The tyranny is not locking down and not using the vaccine to appease the crazies who think it's oppression.	The vaccine mandate is not oppression because it will help to end this pandemic	Liberty/ Oppression

VAD		
Tweet	**Aspect Span**	**Aspect Group**
Study reports Oxford/AstraZeneca vaccine is protective against Brazilian P1 strain of COVID19.	Oxford/AstraZeneca vaccine is protective against Brazilian P1 strain of COVID19	Immunity Level
@user @user @user team, told Reuters while the government admits, it is unknown whether COVID19 mRNA Vaccine BNT162b2 has an impact on fertility.	COVID19 mRNA Vaccine BNT162b2 has an impact on fertility	(Adverse) Side Effects

Table 7.2: Training examples of CMF and VAD. In CMF, Argumentative Patterns are pre-defined phrases indicating an aspect. In VAD, aspect spans are text subsequence of the annotated tweets.

and tweet-level stance classification. For stance classification, we employ RoBERTa and DeBERTa, and use their averaged embeddings for clustering.

7.4.3 Evaluation Metrics

First, we use Clustering Accuracy (CA) and Normalized Mutual Information (NMI) to evaluate the quality of clusters in line with [244, 266]. NMI is defined as NMI $= \big(2 \times \mathrm{I}(y;\hat{y})\big)/\big(\mathrm{H}(y)+\mathrm{H}(\hat{y})\big)$, where $\mathrm{I}(y;\hat{y})$ denotes the mutual information between the ground-truth labels and the predicted labels, $\mathrm{H}(\cdot)$ denotes their entropy. Then we employ BERTScore [318] to evaluate the performance of models in clustering in the absence of ground-truth cluster labels. BERTScore is a successor of Cosine Similarity [115] that measures the sentence distance by calculating the cross distance between their corresponding word embeddings. We follow Bilal et al. [21] to compute the averaged BERTScore as

$$\mathrm{AvgBS} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{\binom{|G_k|}{2}}\sum_{\substack{i,j\in G_k\\ i<j}}\mathrm{BS}(\mathrm{tweet}_i,\mathrm{tweet}_j), \tag{7.4.1}$$

where $|G_k|$ is the size of the k-th group or cluster. We report the average performance for all the models. As a quantitative evaluation metric for disentanglement, we use the Mean Correlation Coefficient (MCC).

7.4.4 Training Details

We experiment with a pre-trained DeBERTa[8] base model. The hidden size is $d_H = 768$. We set both d_V and $d_K = 768$, and $d_{\mathbf{z}} = 1024$. The learning rate is initialised with $\eta = 3e-5$ and the number of epochs is 10. We use Linear Warmup to enforce the triangular learning rate.

We train the model with two Titan RTX graphics cards on a station of an Intel(R) Xeon(R) W-2245 CPU. The training process takes less than 9 hours, with the inference time under 30 minutes.

7.5 Experimental Results

7.5.1 Clustering-Friendly Representation

Clustering Results We first show the advantages of disentangled representations in clustering. With the representations obtained from SBERT and AutoBot,

[8] `https://huggingface.co/docs/transformers/model_doc/deberta-v2`

we employ k-means to perform clustering. Since the similarity between sentences in SBERT is measured by cosine similarity which is less favorable for k-means algorithm, we also use k-medoids to ensure a fair comparison. The other baseline approaches are run with their default settings. We assign the aspect labels to the predicted clusters with the optimal permutation such that the permutation of $\{1, \ldots, K\}$ yields the highest accuracy score, where K denotes the total number of clusters. For the CMF dataset, we set $K = 7$, and on VAD $K = 8$.

Models	**CMF**			**VAD**		
	CA	NMI	Avg BS	CA	NMI	Avg BS
SBERT-k-means	49.2	47.6	18.2	60.5	58.3	19.2
SBERT-k-medoids	50.8	48.1	18.5	62.1	60.1	19.5
SBERT-AHC	51.7	48.5	18.9	64.4	61.2	20.9
AutoBot-k-means	49.2	47.4	18.5	62.8	60.4	20.1
AutoBot-k-medoids	52.5	49.5	19.5	65.6	62.5	20.7
AutoBot-AHC	52.5	48.5	18.9	63.5	60.8	20.5
DS-C-k-means	50.0	47.7	18.5	63.5	60.5	20.7
DS-C-k-medoids	52.5	48.3	18.8	64.7	61.9	21.3
DS-C-k-AHC	50.8	47.8	18.6	64.4	61.5	21.7
VADet	51.7	47.9	18.0	65.4	61.4	20.7
SCCL	48.3	46.9	18.2	63.2	60.8	19.9
RoBERTa-k-means	35.0	35.2	15.0	45.8	46.6	15.7
DeBERTa-k-means	35.8	37.1	15.2	47.7	47.4	16.2
DOC-k-means	51.7	47.8	18.5	64.2	60.7	20.3
DOC-k-medoids	**54.2**	**51.0**	**20.7**	**66.7**	63.1	21.4
DOC-AHC	52.5	49.1	19.1	**66.7**	**63.6**	**22.8**

Table 7.3: Clustering results. Representation learning models are listed with the affiliated clustering methods.

Table 7.3 lists the performance of baseline methods on all the tasks and datasets. We see consistent improvements across all the evaluation metrics using our proposed DOC. When compared with end-to-end methods (i.e., VADet and SCCL) whose intermediate representations cannot be used to calculate a distance, the disparity depends on DOC's clustering approaches employed. On CMF, VADet outperforms SCCL. But DOC gives superior performance overall regardless of the clustering approaches used, showing the flexibility of the DOC representations. In comparisons against representation learning methods, DOC takes the lead as long as it is attached with competent clustering algorithms. This shows the benefit of clustering with disentangled representations since the clustering algorithm will no longer obfuscate the stance polarities and the aspect categories. DOC achieves higher scores on the VAD dataset compared to CMF, with more prominent improvement over the baselines, which may be credited to the increased size of the dataset. When DOC is evaluated with different clustering algorithms, k-medoids excels on CMF, while AHC outperforms the others on VAD, showing that cosine similarity is more appropriate for distance calculation since the k-means algorithm relies on Euclidean distance.

Models	**VAD → CMF**			**CMF → VAD**		
	CA	NMI	Avg BS	CA	NMI	Avg BS
SBERT-AHC	51.6	49.8	19.3	52.4	50.5	17.9
AutoBot-k-medoids	53.1	50.6	20.1	53.7	51.0	18.1
DS-C-k-medoids	54.1	51.2	20.2	54.9	52.4	19.0
VADet	53.5	50.1	19.6	55.2	52.8	19.3
SCCL	48.6	47.0	18.5	53.6	51.6	18.5
DOC-k-medoids	**55.3**	**51.9**	**21.7**	**56.2**	**53.8**	**19.5**
DOC-AHC	53.5	50.4	19.8	55.8	53.7	19.2

Table 7.4: Cross-dataset evaluation results. Each representation learning model is listed with the most performant clustering method.

Cross-Dataset Evaluation In this context, the most interesting property of clustering-friendly representations is their ability to perform clustering in novel datasets whose categories are unknown in advance. To assess this, we use the models trained on CMF to perform clustering on VAD, and repeat the process vice versa. We specify the number of clusters as 7 and 8, respectively. The alignment between the clustered groups and gold labels is solved by the Hungarian algorithm. Note that direct aspect classification across datasets would not be possible since an accurate mapping between the two sets of classes cannot be established. Table 7.4 reports the performance of cross-dataset clustering. Our metrics of interest are still CA, NMI and averaged BERTScore. All the methods show a performance drop on VAD overall, while the performance on CMF turns out to be a bit higher. DOC-k-medoids achieved competitive results across the datasets, demonstrating that clustering-friendly representations disentangle the opinions and, as a result, can integrate unknown aspects.

Stance Classification We report in Table 7.5 the results of DOC, RoBERTa and DeBERTa. For DOC, we only report DOC-AHC since stance labels are by-products of clustering-friendly representations. We see the DOC performance on CMF close

Models	CMF		VAD	
	Micro F1	Macro F1	Micro F1	Macro F1
RoBERTa	72.3±.5	71.2±.4	76.7±.1	75.9±.1
DeBERTa	**74.0**±.6	**73.5**±.6	77.8±.2	**76.8**±.2
DOC-AHC	73.5±.6	72.7±.6	**78.0**±.2	**76.8**±.2

Table 7.5: Stance classification results.

to that of DeBERTa, and that the improvement on VAD is marginal. This may be attributed to the absence of the swapping operation on $\boldsymbol{z}_s$, and therefore the stance latent vector may contain other semantics or noise. Nevertheless, DOC is still preferred over DeBERTa considering its significant performance gain over DeBERTa on aspect clustering.

Model	CMF		VAD	
	CA	AvgBS	CA	AvgBS
Component				
DOC-k-means	**51.7**	**18.5**	**64.2**	**20.3**
w/o pre-trained LM	43.3	16.2	48.4	16.7
w/o inductive bias	50.0	18.0	62.3	19.2
w/o swapped codes	50.8	17.8	62.8	19.0
Choice of Context Vectors				
`MLP`	**51.7**	**18.5**	**64.2**	**20.3**
`CLS`	50.0	17.6	63.2	19.5
`MEAN`	48.3	17.4	60.7	18.7

Table 7.6: Ablation study on removal of components and choices of context vectors.

Ablations Study We study the effects by taking away components of different functionality in disentanglement, and experiment with different choices of context vectors, i.e., $\boldsymbol{u}_s$ and $\boldsymbol{u}_a$. The results are shown in Table 7.6. We see a significant performance drop without loading the pre-trained weights for the language model. The removal of inductive biases and the swapped autoencoder both hamper the clustering of the model across the metrics. The performance gap is more obvious without the inductive bias, which we attribute to the weaker supervision induced by swapping the latent codes. Ablating choices of context vectors shows the superiority of the MLP strategy. In contrast, the performance of the context vector generated by mean pooling is rather poor. It shows that the context vector produced by mean-pooling can hardly trigger the disentanglement of the hidden semantics.

7.5.2 Evaluation of Disentangled Representations

Quantitative Performance As with the nonlinear ICA community [121], we use Mean Correlation Coefficient (MCC) to quantify the extent to which DOC managed to learn disentangled representations. Here, the Point-Biserial Correlation Coefficient between $dist(z_a, \bar{z}_a^k)$ (i.e., the distance between the aspect vector and the centroid of cluster k) and Y (i.e., the dichotomous variable indicating whether it belongs to or not belongs to group k in ground truth) is chosen to measure the isometry between z_a and k. Notice that we specify $dist$ as Euclidean Distance here. However, isometry does not hinge on the Euclidean Distance, and it could be easily substituted with Cosine Similarity, in which case the mean is no longer the best estimation for the cluster center and would be replaced by the medoid of cluster k. The clustering method would be k-medoids accordingly.

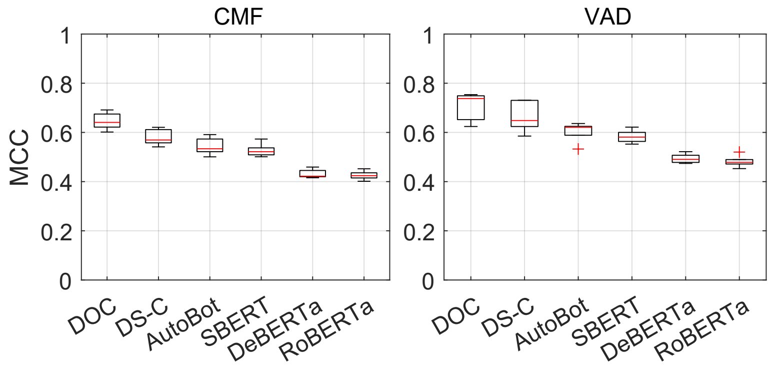


Figure 7.3: Boxplots of MCC for all representation learning models over the 5 runs. The representations are used for k-means clustering in the Euclidean space. A high MCC score indicates the strong correlation between $dist(z_a, \bar{z}_a^k)$ and $z_a \in G_k$.

For each cluster $k \in \{1, 2, \ldots, K\}$, we calculate the correlation coefficient between $dist(z_a, \bar{z}_a^k)$ and Y. We then obtain MCC by averaging the correlation coefficients. A high MCC indicates that the group identity of a data point is closely associated with the geometric position of its z_a in the latent space, which means that z_a captures the group information. The results are shown in Figure 7.3. We observe consistent improvement over the sentence representation models. DS-Clustering is able to encode tweets into aspect embeddings. Nevertheless, its distance between aspect latent vectors is a weaker indicator for group partition compared with that of DOC, suggesting that z_a discovered by DOC better captures the difference between aspects.

Clustering with Different Latent Vectors We experiment clustering using the disentangled aspect vectors $\boldsymbol{z}_a$ or the content vectors $\boldsymbol{z}$ (i.e., without the disentanglement of aspects and stances) on both CMF and VAD datasets, and have the detailed results reported in Table 7.7. It can be observed that using the disentangled aspect vectors for clustering gives better results compared to using the content vectors, regardless of the clustering approaches used. On CMF, the best results are obtained using k-medoids, while on VAD, similar results are obtained using either k-medoids or AHC.

Latent Vector	**CMF**		**VAD**	
	CA	AvgBS	CA	AvgBS
DOC-k-means-z_a	51.7	18.5	64.2	20.3
DOC-k-means-z	48.3	17.5	60.7	18.7
DOC-k-medoids-z_a	**54.2**	**20.7**	**66.7**	21.4
DOC-k-medoids-z	50.8	18.0	61.4	18.9
DOC-AHC-z_a	52.5	19.1	**66.7**	**22.8**
DOC-AHC-z	49.2	17.8	61.9	19.0

Table 7.7: Clustering accuracy and average BERTScore with different latent vectors.

Qualitative Results We illustrate in Figure 7.4 and Figure 7.5 the clustering results and the latent space of the entangled/disentangled representation projected by the t-SNE method. Figure 7.4(a-b) display the cluster assignments after permutation, whereas Figure 7.5(a-b) show the ground-truth labels. The class labels are rendered by colours whose detailed mapping is provided in Figure 7.5. From Figure 7.4, we see clear improvements in terms of clustering quality on both datasets when the model is compared against the DeBERTa-averaged-embedding. Figure 7.5 shows more separated groups thanks to the disentangled representation, providing strong distance-based discrimination for the clustering algorithms. As a result, simple clustering methods like k-means can achieve competitive results against deep clustering methods (i.e., SCCL and VAD), which have access to weak labels or data augmentations.

Color Mappings in Visualisation We illustrate in Figure 7.5 the color mapping from t-SNE plots to the true aspect category labels. It is shown that the vectors are more separated and their grouping aligns closer to the ground-truth labels when they are clustered on the space of z_a, indicating that such latent vectors provide strong distance-based discrimination among groups in the Euclidean space, as has

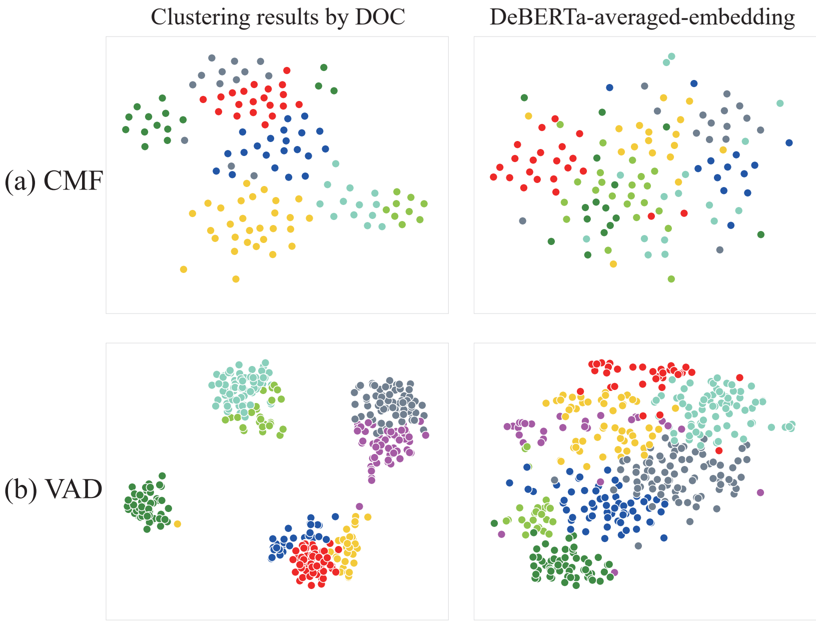


Figure 7.4: 2-D plots of the data points projected by t-SNE.

been used as a distance metric in the t-SNE algorithm. We also experiment with cosine-similarity metric for k-medoids and the results have been reported in the Experiments section.

7.6 Limitations

There are a few limitations we would like to address. First of all, the number of clusters needs manual configuration. This is a limitation of the clustering algorithms [299] since we need to set a threshold for convergence, which consequentially pinpoints k. An expedient alternative is to analyse the dataset for the realistic settings or probe into k for the optimal setup, which is, however, beyond the scope of this work. Another limitation is the pre-requisite for millions of unannotated data. The autoencoder needs enormous data to learn bottleneck representations. Its performance would be hindered without access to abundant corpora. Lastly, the performance of the acquired clustering-friendly representations depends on the similarity metric chosen. Efforts need to be made to find the best option, whether it is Euclidean distance or cosine similarity, etc.

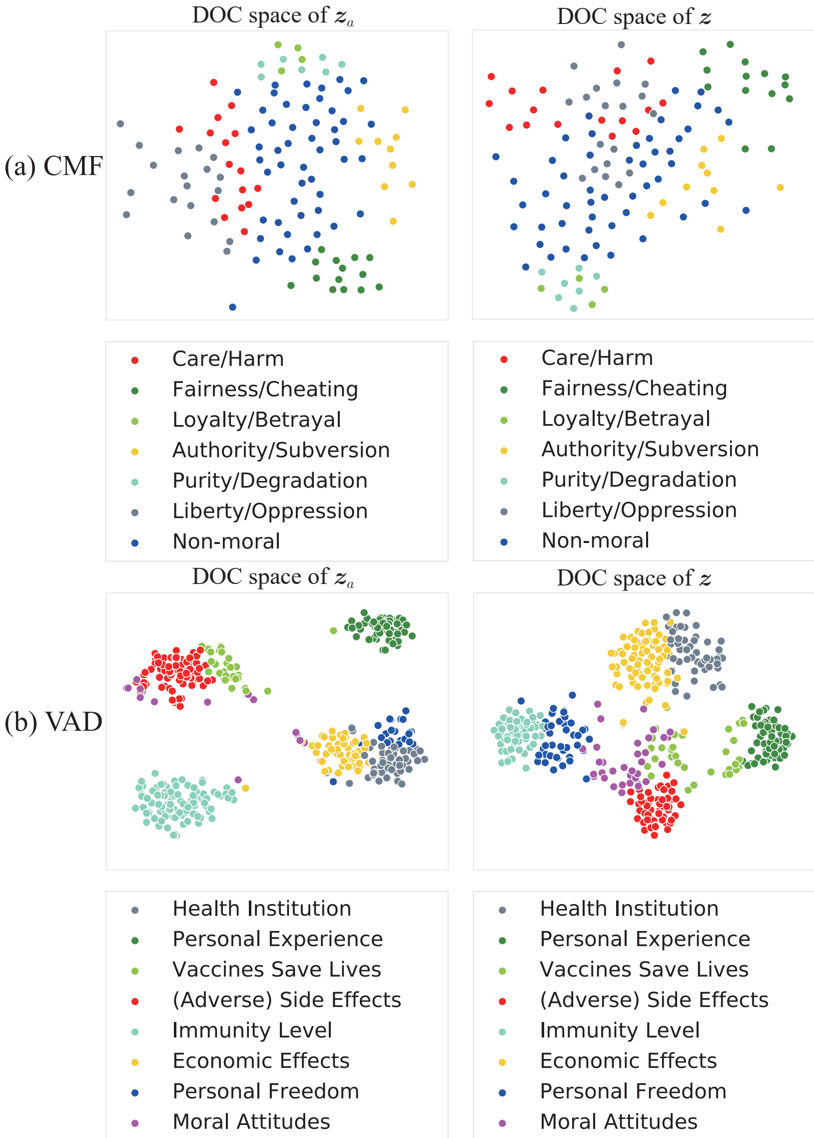


Figure 7.5: t-SNE plots on CMF and VAD. Each dot is a tweet encoded using either the disentangled aspect vector z_a (left subfigure) or the latent content vector z (right subfigure). Different colors indicate the true aspect category labels.

7.7 Summary

We have introduced DOC, a *Disentangled Opinion Clustering* model for vaccination opinion mining from social media. DOC is able to disentangle users' stances from opinions via a disentangling attention mechanism and a swap-autoencoder. It was designed to process unseen aspect categories thanks to the clustering approach, leveraging *clustering-friendly* representations induced by out-of-the-box Sentence-BERT encodings and the disentangling mechanisms. The experimental assessment demonstrated the benefit of the disentangling mechanism on the quality of aspect-based clusters and the generalization capability across datasets with different aspect categories outperforming existing approaches in terms of generalisation and coherence of the generated clusters.

Chapter 8

Conclusion

In this thesis, we have worked on topic representation learning on sequential data, with applications to text classification and clustering.

The thesis addresses several limitations in sequence-to-sequence modelling, as explained in §1.1 and §2.2, to increase the machines' understanding of the text. We develop methods for word representation learning and disentanglement-focused sentence-level representation learning. Topic representation learning is also delved into for capturing the semantics holistically and for the generalization across datasets. We show the benefits of using topic representations and pre-trained word representations in sequence-to-sequence modelling by testing the proposed approaches on several classification tasks. The benefits of topic modelling and sequence-to-sequence fine-tuning also include increased generalization, as reflected by text clustering results.

The models we designed are combinations of autoregressive models and latent variable models according to the taxonomy elaborated in §2.2. In particular, the autoregressive models are pre-trained LMs designated for sequence-to-sequence prediction, which are advantageous in gauging dependencies within local contexts. There are also sequence-to-sequence modules (e.g., Transformers) employed for higher-level dependencies, such as documents or conversations. The employment of latent variable models leads the model to compress the collocations into fixed-size representations, allowing the interpolation over the semantic space and disentanglement of latent variables, thus leading to increased generalization.

In what follows, we summarize the contributions with regard to each research question, pointing out the limitations while outlooking the research directions to the future.

8.1 Overall Summary

Having introduced the motivation and listed the contributions in §1, we discuss the related literature in §2 in a taxonomy where each branch is chronologically updated. We then present the main body of our work in §3 - 7. Now, we attempt to summarize the approaches to the research objectives outlined in §1.2.

To model the intricate dependencies between different levels of text, we design the Joint Topic Word-embedding (JTW) model to gauge semantics at the word level and sentence level simultaneously (§3). We develop a neural opinion dynamics model, called NTOM, to forecast users' stances in their timelines based on sequence-to-sequence prediction (§4). Chapter §5 utilises Transformers to detect emotions of utterances in dialogues, and Chapter §6 - §7 employs pre-trained language models to provide text span prediction or learn inductive biases from a text sequence.

The research question of how to capture holistic properties of sequential data is answered by topic representation learning methods developed in Chapter §3 where topics are explicitly modelled as a matrix, and Chapter §5 - 7 where topics are attained with fine-tuning of language models. We insert a latent variable into hidden layers as a bottleneck representation and form the topic representation learning as denoising an LM-based auto-encoder.

The VADet model (§6) provides a semi-supervised framework that can acquire topic representations via denoising auto-encoder and disentanglement of such representations can be achieved in fine-tuning with constrained priors and inductive biases. Additionally, the DOC model (§7) induces the disentanglement by disentangled cross attention and swapping auto-encoder. We find that disentangled learning is promising in detecting aspect-related text spans. The disentangled representations are clustering-friendly with distance metrics, which allow for improved flexibility across a range of applied datasets.

We provide extensive analysis of the quality of the acquired representations. In particular, the neural sequence models (i.e., NTOM, TodKAT and VADet) are tested by sequence labelling on social media or conversational datasets. The improved performance shows that sequence modelling is a viable approach. Meanwhile, the quality of topic representations is evaluated from perspectives of clustering and usefulness in classification. Hence, the research question can be answered affirmatively.

We build a dataset, namely the VAD dataset (§6.4.1), from social media text and annotate the dataset with aspect text spans and stance labels, which allows for the evaluation of attitude detection. The annotation contains a categorical label indicating the aspect category for the evaluation of clustering and latent topic disentanglement. The dataset is supplemented with a large unannotated corpus to learn the topic representation before the supervised learning.

8.2 Limitations and Outlooks

Chapter §3 studies the joint topic word embedding model. One limitation is on pivot words of the sliding window, which are presumably independent. It is more realistic to introduce dependencies between pivot words as implied by Bamler and Mandt [12], in which circumstance the context scope will be implicitly expanded to the entire document. The discourse relationships can also be considered to model the semantic drifts between different contexts. Amid the recent development of LLMs that encapsulate the word representations and provide the standardised outputs as free-form text or multiple-choice selection, it is desirable to learn topic representations which the standardised predictions can be grounded into. Such topic representations can sit in the middle, finding supporting examples with semantic search.

Chapter §4 models the social impacts with a fixed-size neighbourhood context. It is, however, possible to use attention-based aggregation (e.g., Graph Attention Nets [279] and Variational Graph Auto-Encoders [126]) to account for heterogeneous structure and contents. It is also feasible to encode the graph structure into hyperbolic representations. For the usage of LLMs in social media analysis, how to simulate social skills remains an open problem [168]. Work needs to be done to apply LLMs to tasks that require structure prediction. A user case is to instruct an LLM to generate the graph structure in a layout language.

The topic representation learning approach presented in Chapter §5 - 7 frames the topics as intermediate latent variables between LM hidden layers, based on the notion that different levels' hidden states capture different abstract levels. However, it could also be helpful to consider multiple layers of latent variables (e.g., the diffusion process) for richer representations. The disentanglement of the latent variable is induced by a factorized and conditional prior in VAD, and for DOC the inductive bias is the isometry between the latent code and the group label. However, the assumption that the disentangled latent codes align with the desired factors does not always hold (e.g., in some cases the aspects are biased and thus independent of the stance). In such circumstances, latent relations or combinatorial structures should be considered. This could naturally be the next step in this vein.

8.3 Future Directions

Large Language Models have unified a wide range of NLP tasks (§ 2.4.2) recently. In lieu of the modular nature of GPT-4 [198] which combines diverse objectives for pre-training and employs RLHF for instruction-tuning, it is natural to consider wrapping the foundation models with a general reinforcement learning algorithm that masters chess and other games [251], or automata theories that encompass Turing Machines [274]. We can draw an analogy between the foundation models and cortex of brain that extracts representations or meanings of words. These representation extractors and knowledge retrievers or memory indexers are synergically operated by brain moderators in this sense. For topic representation learning here, modules can be built to steer LLMs to produce those low-dimensional, clustering-friendly representations. From the dataset perspective, there is surge of need for acquiring diverse prompts, besides a multitude of training examples formed as prompt-completion pairs [199]. However, the supervision provided by the data points is limited, since the reward is implicit and the model needs to extrapolate from the annotations, which has been implemented by reinforcement learning from human feedbacks. In contrast, the open world [222] presents an ideal source of supervision, where the model can be rewarded with incentives derived from world mechanisms [9] that synergistically complements the audience model with the everyday commonsense.

Bibliography

[1] Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008. URL https://link.springer.com/book/10.1007/978-0-387-68560-1.

[2] Rilwan Adewoyin, Ritabrata Dutta, and Yulan He. RSTGen: Imbuing fine-grained interpretable control into long-FormText generators. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1822–1835, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.133. URL https://aclanthology.org/2022.naacl-main.133.

[3] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken ELBO. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 159–168. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/alemi18a.html.

[4] Hamidreza Alvari and Paulo Shakarian. Hawkes process for understanding the influence of pathogenic social media accounts. In *2nd International Conference on Data Intelligence and Security, ICDIS 2019, South Padre Island, TX, USA, June 28-30, 2019*, pages 36–42. IEEE, 2019. doi: 10.1109/ICDIS.2019.00013. URL https://doi.org/10.1109/ICDIS.2019.00013.

[5] Nicholas Andrews and Marcus Bishop. Learning invariant representations of social media users. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1684–1695, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1178. URL https://aclanthology.org/D19-1178.

[6] Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23284–23296. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/93be245fce00a9bb2333c17ceae4b732-Paper-Conference.pdf.

[7] Meysam Asgari-Chenaghlu, Mohammad-Reza Feizi-Derakhshi, Mohammad-Ali Balafar, Cina Motamed, et al. Topicbert: A transformer transfer learning based memory-graph approach for multimodal streaming social media topic detection. *arXiv preprint arXiv:2008.06877*, 2020. URL https://arxiv.org/abs/2008.06877.

[8] Christopher L. Asplund, J. Jay Todd, Andy P. Snyder, and René Marois. A central role for the lateral prefrontal cortex in goal-directed and stimulus-driven attention. *Nature Neuroscience*, 13(4):507–512, Apr 2010. ISSN 1546-1726. doi: 10.1038/nn.2509. URL https://doi.org/10.1038/nn.2509.

[9] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, Vancouver, Canada, June 2023. IEEE Computer Society. URL https://openaccess.thecvf.com/content/CVPR2023/papers/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.pdf.

[10] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

[12] Robert Bamler and Stephan Mandt. Dynamic word embeddings. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Re-*

search, pages 380–389. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/bamler17a.html`.

[13] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. A large-scale covid-19 twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia*, 2(3):315–324, 2021. ISSN 2673-3986. doi: 10.3390/epidemiologia2030024. URL `https://www.mdpi.com/2673-3986/2/3/24`.

[14] Yamini Bansal, Gal Kaplun, and Boaz Barak. For self-supervised learning, rationality implies generalization, provably. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=Srmggo3b3X6`.

[15] Oren Barkan. Bayesian neural word embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017. doi: 10.1609/aaai.v31i1.10987. URL `https://ojs.aaai.org/index.php/AAAI/article/view/10987`.

[16] Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. Structured sentiment analysis as dependency graph parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3387–3402, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.263. URL `https://aclanthology.org/2021.acl-long.263`.

[17] Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2126. URL `https://aclanthology.org/S17-2126`.

[18] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003. URL `http://www.jmlr.org/papers/v3/bengio03a.html`.

[19] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50. URL `https://doi.org/10.1109/TPAMI.2013.50`.

[20] Federico Bianchi, Silvia Terragni, and Dirk Hovy. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.96. URL `https://aclanthology.org/2021.acl-short.96`.

[21] Iman Munire Bilal, Bo Wang, Maria Liakata, Rob Procter, and Adam Tsakalidis. Evaluation of thematic coherence in microblogs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 6800–6814. Association for Computational Linguistics, August 2021. doi: 10.18653/v1/2021.acl-long.530. URL `https://aclanthology.org/2021.acl-long.530`.

[22] David M. Blei and Peter I. Frazier. Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12(74):2461–2488, 2011. URL `http://jmlr.org/papers/v12/blei11a.html`.

[23] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of machine Leraning research*, 3:993–1022, March 2003. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=944919.944937`.

[24] David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2), feb 2010. ISSN 0004-5411. doi: 10.1145/1667053.1667056. URL `https://doi.org/10.1145/1667053.1667056`.

[25] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL `https://doi.org/10.1080/01621459.2017.1285773`.

[26] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows,

energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7327–7347, 2022. doi: 10.1109/TPAMI.2021.3116668. URL `https://ieeexplore.ieee.org/document/9555209`.

[27] Tom Bosc and Pascal Vincent. Do sequence-to-sequence VAEs learn global features of sentences? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4296–4318, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.350. URL `https://aclanthology.org/2020.emnlp-main.350`.

[28] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1470. URL `https://aclanthology.org/P19-1470`.

[29] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11867. URL `https://ojs.aaai.org/index.php/AAAI/article/view/11867`.

[30] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL `https://aclanthology.org/D15-1075`.

[31] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1002. URL `https://aclanthology.org/K16-1002`.

[32] Arthur Bražinskas, Serhii Havrylov, and Ivan Titov. Embedding words as distributions with a Bayesian skip-gram model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1775–1789, Santa

Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1151.

[33] Eleftheria Briakou, Nikos Athanasiou, and Alexandros Potamianos. Cross-topic distributional semantic representations via unsupervised mappings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1052–1061, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1110. URL https://aclanthology.org/N19-1110.

[34] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[35] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018. URL https://arxiv.org/abs/1804.03599.

[36] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008. URL https://link.springer.com/article/10.1007/s10579-008-9076-6.

[37] Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy, July 2019.

Association for Computational Linguistics. doi: 10.18653/v1/P19-1563. URL https://aclanthology.org/P19-1563.

[38] Ranganathan Chandrasekaran, Rashi Desai, Harsh Shah, Vivek Kumar, and Evangelos Moustakas. Examining public sentiments and attitudes toward covid-19 vaccination: Infoveillance study using twitter posts. *JMIR Infodemiology*, 2(1):e33909, Apr 2022. ISSN 2564-1891. doi: 10.2196/33909. URL https://infodemiology.jmir.org/2022/1/e33909.

[39] Yatin Chaudhary, Pankaj Gupta, Khushbu Saxena, Vivek Kulkarni, Thomas Runkler, and Hinrich Schütze. TopicBERT for energy efficient document classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1682–1690, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.152. URL https://aclanthology.org/2020.findings-emnlp.152.

[40] Chengyao Chen, Zhitao Wang, Yu Lei, and Wenjie Li. Content-based influence modeling for opinion behavior prediction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2207–2216, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1208.

[41] Chengyao Chen, Zhitao Wang, and Wenjie Li. Tracking dynamics of opinion behaviors with a content-based sequential opinion influence model. *IEEE Transactions on Affective Computing*, 11(4):627–639, 2020. doi: 10.1109/TAFFC.2018.2821123. URL https://ieeexplore.ieee.org/document/8328845?utm_source=researcher_app&utm_medium=referral&utm_campaign=RESR_MRKT_Researcher_inbound.

[42] Henry Y. Chen, Ethan Zhou, and Jinho D. Choi. Robust coreference resolution and entity linking on dialogues: Character identification on TV show transcripts. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 216–225, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-1023. URL https://aclanthology.org/K17-1023.

[43] Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett,

editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/1ee3dfcd8a0645a25a35977997223d22-Paper.pdf.

[44] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1053. URL https://aclanthology.org/D16-1053.

[45] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. URL https://arxiv.org/abs/1904.10509.

[46] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL https://aclanthology.org/W14-4012.

[47] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179.

[48] Noam Chomsky. *On Nature and Language*. Cambridge University Press, 2002. doi: 10.1017/CBO9780511613876. URL https://www.cambridge.org/core/books/on-nature-and-language/210D36D24C9F0397E7599D6AB5F2ACB1.

[49] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking attention with performers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

[50] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. URL `https://arxiv.org/abs/2204.02311`.

[51] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL `https://proceedings.neurips.cc/paper/2015/file/b618c3210e934362ac261db280128c22-Paper.pdf`.

[52] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1078. URL `https://aclanthology.org/P18-1078`.

[53] Shuyang Dai, Zhe Gan, Yu Cheng, Chenyang Tao, Lawrence Carin, and Jingjing Liu. APo-VAE: Text generation in hyperbolic space. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 416–431, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.36. URL `https://aclanthology.org/2021.naacl-main.36`.

[54] Siddharth Dalmia, Brian Yan, Vikas Raunak, Florian Metze, and Shinji Watanabe. Searchable hidden intermediates for end-to-end models of decomposable sequence tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1882–1896, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.151. URL `https://aclanthology.org/2021.naacl-main.151`.

[55] Rajarshi Das, Manzil Zaheer, and Chris Dyer. Gaussian LDA for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 795–

804, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1077. URL `https://aclanthology.org/P15-1077`.

[56] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=H1edEyBKDS`.

[57] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=H1edEyBKDS`.

[58] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990. doi: https://doi.org/10.1002/(SICI)1097-4571(199009)41:6⟨391::AID-ASI1⟩3.0.CO;2-9. URL `https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9`.

[59] Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández. You shall know a user by the company it keeps: Dynamic representations for social media users in NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4707–4717, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1477. URL `https://aclanthology.org/D19-1477`.

[60] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota,

June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

[61] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL `http://arxiv.org/abs/1410.8516`.

[62] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2793–2803. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/dong21a.html`.

[63] John Dougrez-Lewis, Maria Liakata, Elena Kochkina, and Yulan He. Learning disentangled latent topics for Twitter rumour veracity classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3902–3908, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.341. URL `https://aclanthology.org/2021.findings-acl.341`.

[64] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1555–1564, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939875. URL `https://doi.org/10.1145/2939672.2939875`.

[65] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4): 384, 1993. URL `https://psycnet.apa.org/record/1993-32252-001`.

[66] Le Fang, Chunyuan Li, Jianfeng Gao, Wen Dong, and Changyou Chen. Implicit deep latent variable models for text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3946–3956, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1407. URL `https://aclanthology.org/D19-1407`.

[67] William Fedus, Ian J. Goodfellow, and Andrew M. Dai. Maskgan: Better text generation via filling in the _______. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL `https://openreview.net/forum?id=ByOExmWAb`.

[68] Lev Finkelstein, Evgeniy Gabrilovich1, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131, jan 2002. ISSN 1046-8188. doi: 10.1145/503104.503110. URL `https://doi.org/10.1145/503104.503110`.

[69] James Foulds. Mixed membership word embeddings for computational social science. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 86–95. PMLR, 09–11 Apr 2018. URL `https://proceedings.mlr.press/v84/foulds18a.html`.

[70] Andrea Galassi, Marco Lippi, and Paolo Torroni. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4291–4308, 2021. doi: 10.1109/TNNLS.2020.3019893. URL `https://ieeexplore.ieee.org/document/9194070`.

[71] Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, Guodong Zhou, and Shuming Shi. A discrete CVAE for response generation on short-text conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1898–1908, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1198. URL `https://aclanthology.org/D19-1198`.

[72] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/gehring17a.html`.

[73] Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*,

85(410):398–409, 1990. doi: 10.1080/01621459.1990.10476213. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10476213.

[74] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984. doi: 10.1109/TPAMI.1984.4767596. URL https://ieeexplore.ieee.org/document/4767596.

[75] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1015. URL https://aclanthology.org/D19-1015.

[76] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.224. URL https://aclanthology.org/2020.findings-emnlp.224.

[77] Yichen Gong, Heng Luo, and Jian Zhang. Natural language inference over interaction space. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1dHXnH6-.

[78] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[79] Koustava Goswami, Rajdeep Sarkar, Bharathi Raja Chakravarthi, Theodorus Fransen, and John P. McCrae. Unsupervised deep language and dialect identification for short texts. In *Proceedings of the 28th International Conference*

on Computational Linguistics, pages 1606–1617, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.141. URL https://aclanthology.org/2020.coling-main.141.

[80] Edouard Grave, Moustapha M Cisse, and Armand Joulin. Unbounded cache model for online language modeling with open vocabulary. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/f44ee263952e65b3610b8ba51229d1f9-Paper.pdf.

[81] Alex Graves and Jürgen Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/file/66368270ffd51418ec58bd793f2d9b1b-Paper.pdf.

[82] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013. URL https://ieeexplore.ieee.org/document/6638947.

[83] Alexander Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*. PhD thesis, Technische Universität München, 2008. URL https://www.cs.toronto.edu/~graves/phd.pdf.

[84] Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. Hierarchical topic models and the nested chinese restaurant process. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper/2003/file/7b41bfa5085806dfa24b8c9de0ce567f-Paper.pdf.

[85] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004. doi: 10.1073/pnas.0307752101. URL http://www.pnas.org/content/101/suppl_1/5228.abstract.

[86] Yosh Halberstam and Brian Knight. Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. *Journal*

of Public Economics, 143:73–88, 2016. ISSN 0047-2727. doi: https://doi.org/10.1016/j.jpubeco.2016.08.011. URL https://www.sciencedirect.com/science/article/pii/S0047272716301001.

[87] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

[88] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 1483–1492. IEEE, 2019. doi: 10.1109/ICCVW.2019.00186. URL https://doi.org/10.1109/ICCVW.2019.00186.

[89] Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971. ISSN 00063444. URL http://www.jstor.org/stable/2334319.

[90] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf.

[91] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

[92] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1048. URL https://aclanthology.org/P19-1048.

[93] Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4246–4253, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.379. URL https://aclanthology.org/2020.findings-emnlp.379.

[94] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

[95] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. URL http://arxiv.org/abs/1812.02230.

[96] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

[97] Daniella Horan, Eitan Richardson, and Yair Weiss. When is unsupervised disentanglement possible? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 5150–5161. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/29586cb449c90e249f1f09a0a4ee245a-Paper.pdf.

[98] Christian Horvat and Jean-Pascal Pfister. Denoising normalizing flow. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9099–9111. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/4c07fe24771249c343e70c32289c1192-Paper.pdf.

[99] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339,

Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031. URL https://aclanthology.org/P18-1031.

[100] Chao-Chun Hsu and Lun-Wei Ku. SocialNLP 2018 EmotionX challenge overview: Recognizing emotions in dialogues. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 27–31, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-3505. URL https://aclanthology.org/W18-3505.

[101] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/0a0a0c8aaa00ade50f74a3f0ca981ed7-Paper.pdf.

[102] Tao Hu, Siqin Wang, Wei Luo, Mengxi Zhang, Xiao Huang, Yingwei Yan, Regina Liu, Kelly Ly, Viraj Kacker, Bing She, and Zhenlong Li. Revealing public opinion towards covid-19 vaccines with twitter data in the united states: Spatiotemporal perspective. *J Med Internet Res*, 23(9):e30854, Sep 2021. ISSN 1438-8871. doi: 10.2196/30854. URL https://www.jmir.org/2021/9/e30854.

[103] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Toward controlled generation of text. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/hu17e.html.

[104] Jiaxin Huang, Yu Meng, Fang Guo, Heng Ji, and Jiawei Han. Weakly-supervised aspect-based sentiment analysis via joint aspect-sentiment topic embedding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6989–6999, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.568. URL https://aclanthology.org/2020.emnlp-main.568.

[105] Amir Hussain, Ahsen Tahir, Zain Hussain, Zakariya Sheikh, Mandar Gogate, Kia Dashtipour, Azhar Ali, and Aziz Sheikh. Artificial intelligence–enabled

analysis of public attitudes on facebook and twitter toward covid-19 vaccines in the united kingdom and the united states: Observational study. *J Med Internet Res*, 23(4):e26627, Apr 2021. ISSN 1438-8871. doi: 10.2196/26627. URL https://www.jmir.org/2021/4/e26627.

[106] Amir Hussain, Ahsen Tahir, Zain Hussain, Zakariya Sheikh, Mandar Gogate, Kia Dashtipour, Azhar Ali, and Aziz Sheikh. Artificial intelligence–enabled analysis of public attitudes on facebook and twitter toward covid-19 vaccines in the united kingdom and the united states: Observational study. *J Med Internet Res*, 23(4):e26627, Apr 2021. ISSN 1438-8871. doi: 10.2196/26627. URL https://www.jmir.org/2021/4/e26627.

[107] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/d305281faf947ca7acade9ad5c8c818c-Paper.pdf.

[108] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 859–868. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/hyvarinen19a.html.

[109] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430, 2000. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(00)00026-5. URL https://www.sciencedirect.com/science/article/pii/S0893608000000265.

[110] Ignacio Iacobacci and Roberto Navigli. LSTMEmbed: Learning word and sense representations from a large semantically annotated corpus with long short-term memories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1685–1695, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1165. URL https://aclanthology.org/P19-1165.

[111] Shoaib Jameel and Steven Schockaert. Word and document embedding with vMF-mixture priors on context word vectors. In *Proceedings of the 57th An-*

nual Meeting of the Association for Computational Linguistics*, pages 3319–3328, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1321. URL https://aclanthology.org/P19-1321.

[112] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

[113] Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R. Lyu. HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 397–406, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1037. URL https://aclanthology.org/N19-1037.

[114] Wenxiang Jiao, Michael Lyu, and Irwin King. Real-time emotion recognition via attention gated hierarchical memory network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8002–8009, Apr. 2020. doi: 10.1609/aaai.v34i05.6309. URL https://ojs.aaai.org/index.php/AAAI/article/view/6309.

[115] Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1041. URL https://aclanthology.org/P19-1041.

[116] Josef Jon, João Paulo Aires, Dusan Varis, and Ondřej Bojar. End-to-end lexically constrained machine translation for morphologically rich languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4019–4033, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.311. URL https://aclanthology.org/2021.acl-long.311.

[117] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. *An Introduction to Variational Methods for Graphical Models*, page

105–161. MIT Press, Cambridge, MA, USA, 1999. ISBN 0262600323. URL https://dl.acm.org/doi/10.5555/308574.308660.

[118] Tom Joy, Sebastian M. Schmon, Philip H. S. Torr, Siddharth Narayanaswamy, and Tom Rainforth. Capturing label characteristics in vaes. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=wQRlSUZ5V7B.

[119] Biing Hwang Juang and Laurence R Rabiner. Hidden markov models for speech recognition. *Technometrics*, 33(3):251–272, 2023/07/21/ 1991. ISSN 00401706. doi: 10.2307/1268779. URL http://www.jstor.org/stable/1268779. Full publication date: Aug., 1991.

[120] Andrej Karpathy, Pieter Abbeel, Greg Brockman, Peter Chen, Vicki Cheung, Rocky Duan, Ian Goodfellow, Durk Kingma, Jonathan Ho, Rein Houthooft, Tim Salimans, John Schulman, Ilya Sutskever, and Wojciech Zaremba. Generative models. *OpenAI*, 2016. URL https://openai.com/blog/generative-models/.

[121] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

[122] Yoon Kim, Sam Wiseman, and Alexander M. Rush. A tutorial on deep latent variable models of natural language, 2018. URL https://arxiv.org/abs/1812.06834.

[123] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

[124] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in*

Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/d139db6a236200b21cc7f752979132d0-Paper.pdf.

[125] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/d523773c6b194f37b938d340d5d02232-Paper.pdf.

[126] Thomas N. Kipf and Max Welling. Variational graph auto-encoders. In *NeurIPS Workshop on Bayesian Deep Learning*, 2016. URL http://arxiv.org/abs/1611.07308.

[127] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.* OpenReview.net, 2020. URL https://openreview.net/forum?id=rkgNKkHtvB.

[128] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.* OpenReview.net, 2020. URL https://openreview.net/forum?id=rkgNKkHtvB.

[129] Ryota Kobayashi and Renaud Lambiotte. Tideh: Time-dependent hawkes process for predicting retweet dynamics. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):191–200, Aug. 2021. doi: 10.1609/icwsm.v10i1.14717. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14717.

[130] Florian Kunneman, Mattijs Lambooij, Albert Wong, Antal Van Den Bosch, and Liesbeth Mollema. Monitoring stance towards vaccination in twitter messages. *BMC medical informatics and decision making*, 20(1):1–14, 2020. URL https://doi.org/10.1186/s12911-020-1046-y.

[131] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June 2016. Association for

Computational Linguistics. doi: 10.18653/v1/N16-1030. URL https://aclanthology.org/N16-1030.

[132] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

[133] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. URL https://yann.lecun.com/exdb/publis/pdf/lecun-06.pdf.

[134] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. URL https://www.nature.com/articles/nature14539.

[135] Kalev Leetaru and Philip A. Schrodt. Gdelt: Global data on events, location, and tone. *ISA Annual Convention*, 2013. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.686.6605.

[136] Kaufman Leonard and J Rousseeuw Peter. Finding groups in data: an introduction to cluster analysis. *Probability and Mathematical Statistics. Applied Probability and Statistics*, 1990. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316801.

[137] Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. {PMI}-masking: Principled masking of correlated spans. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3Aoft6NWFej.

[138] Chang Li and Dan Goldwasser. Encoding social information with graph convolutional networks forPolitical perspective detection in news media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2594–2604, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1247. URL https://aclanthology.org/P19-1247.

[139] Chenliang Li, Yu Duan, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Trans. Inf. Syst.*, 36(2), aug 2017. ISSN 1046-8188. doi: 10.1145/3091108. URL https://doi.org/10.1145/3091108.

[140] Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.378. URL `https://aclanthology.org/2020.emnlp-main.378`.

[141] Hongjing Li, Hanqi Yan, Yanran Li, Li Qian, Yulan He, and Lin Gui. Distinguishability calibration to in-context learning. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1385–1397, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL `https://aclanthology.org/2023.findings-eacl.102`.

[142] Jiazheng Li, Zhaoyue Sun, Bin Liang, Lin Gui, and Yulan He. CUE: An uncertainty interpretation framework for text classifiers built on pre-trained language models. In Robin J. Evans and Ilya Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1253–1262. PMLR, 31 Jul–04 Aug 2023. URL `https://proceedings.mlr.press/v216/li23d.html`.

[143] Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1865–1874, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1169. URL `https://aclanthology.org/N18-1169`.

[144] Wei Li, Wei Shao, Shaoxiong Ji, and Erik Cambria. Bieru: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing*, 467:73–82, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.09.057. URL `https://www.sciencedirect.com/science/article/pii/S0925231221014351`.

[145] Xin Li, Lidong Bing, Wai Lam, and Bei Shi. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1087. URL `https://aclanthology.org/P18-1087`.

[146] Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. Aspect term extraction with history attention and selective transformation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4194–4200. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/583. URL `https://doi.org/10.24963/ijcai.2018/583`.

[147] Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5505. URL `https://aclanthology.org/D19-5505`.

[148] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL `https://aclanthology.org/I17-1099`.

[149] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. Gated graph sequence neural networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL `http://arxiv.org/abs/1511.05493`.

[150] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 375–384, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585123. doi: 10.1145/1645953.1646003. URL `https://doi.org/10.1145/1645953.1646003`.

[151] Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. Heterogeneous graph attention networks for semi-supervised short text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4821–4830, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1488. URL `https://aclanthology.org/D19-1488`.

[152] Jun S. Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994. ISSN 01621459. URL http://www.jstor.org/stable/2290921.

[153] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 276–284, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/liub16.html.

[154] Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Topical word embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015. doi: 10.1609/aaai.v29i1.9522. URL https://ojs.aaai.org/index.php/AAAI/article/view/9522.

[155] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

[156] Micha Livne, Kevin Swersky, and David J. Fleet. Sentencemim: A latent variable language model, 2020. URL https://arxiv.org/abs/2003.02645.

[157] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

[158] Francesco Locatello, Ben Poole, Gunnar Raetsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6348–6359. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/locatello20a.html.

[159] Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. Disentangling factors of variations using few labels. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=SygagpEKwB`.

[160] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/9dcb88e0137649590b755372b040afad-Paper.pdf`.

[161] Junru Lu, Gabriele Pergola, Lin Gui, Binyang Li, and Yulan He. CHIME: Cross-passage hierarchical memory network for generative review question answering. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2547–2560, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.229. URL `https://aclanthology.org/2020.coling-main.229`.

[162] Junru Lu, Xingwei Tan, Gabriele Pergola, Lin Gui, and Yulan He. Event-centric question answering via contrastive learning and invertible event transformation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2377–2389, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.findings-emnlp.176`.

[163] Michal Lukasik, P. K. Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 393–398, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2064. URL `https://aclanthology.org/P16-2064`.

[164] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Lin-

guistics. doi: 10.18653/v1/D15-1166. URL https://aclanthology.org/D15-1166.

[165] Joanne Chen Lyu, Eileen Le Han, and Garving K Luli. Covid-19 vaccine–related discussion on twitter: Topic modeling and sentiment analysis. *J Med Internet Res*, 23(6):e24435, Jun 2021. ISSN 1438-8871. doi: 10.2196/24435. URL https://www.jmir.org/2021/6/e24435.

[166] Yukun Ma, Haiyun Peng, and Erik Cambria. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/12048.

[167] Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. Variational information bottleneck for effective low-resource fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=kvhzKz-_DMF.

[168] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*, 2023. URL https://arxiv.org/abs/2301.06627.

[169] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825, Jul. 2019. doi: 10.1609/aaai.v33i01.33016818. URL https://ojs.aaai.org/index.php/AAAI/article/view/4657.

[170] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. *CoRR*, abs/1511.05644, 2015. URL http://arxiv.org/abs/1511.05644.

[171] Diego Marcheggiani, Oscar Täckström, Andrea Esuli, and Fabrizio Sebastiani. Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann, editors, *Advances in Information Retrieval*, pages 273–285, Cham, 2014.

Springer International Publishing. ISBN 978-3-319-06028-6. URL https://link.springer.com/chapter/10.1007/978-3-319-06028-6_23#citeas.

[172] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/ef0917ea498b1665ad6c701057155abe-Paper.pdf.

[173] Diana McCarthy and Roberto Navigli. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/S07-1009.

[174] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001. ISSN 03600572, 15452115. URL http://www.jstor.org/stable/2678628.

[175] Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/6463c88460bd63bbe256e495c63aa40b-Paper.pdf.

[176] Oren Melamud, Omer Levy, and Ido Dagan. A simple word embedding model for lexical substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 1–7, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-1501. URL https://aclanthology.org/W15-1501.

[177] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. Weakly-supervised hierarchical text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6826–6833, Jul. 2019. doi: 10.1609/aaai.v33i01.33016826. URL https://ojs.aaai.org/index.php/AAAI/article/view/4658.

[178] Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. Topic discovery via latent space clustering of pretrained language model repre-

sentations. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3143–3152, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3512034. URL https://doi.org/10.1145/3485447.3512034.

[179] Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1727–1736, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/miao16.html.

[180] Tomás Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, editors, *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048. ISCA, 2010. URL http://www.isca-speech.org/archive/interspeech_2010/i10_1045.html.

[181] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. URL http://arxiv.org/abs/1301.3781.

[182] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999792.2999959.

[183] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1147. URL https://aclanthology.org/D16-1147.

[184] Sebastião Miranda, Artūrs Znotiņš, Shay B. Cohen, and Guntis Barzdins.

Multilingual clustering of streaming news. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4535–4544, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1483. URL `https://aclanthology.org/D18-1483`.

[185] Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL `https://aclanthology.org/D13-1171`.

[186] Joel Ruben Antony Moniz and David Krueger. Nested lstms. In Min-Ling Zhang and Yung-Kyun Noh, editors, *Proceedings of the Ninth Asian Conference on Machine Learning*, volume 77 of *Proceedings of Machine Learning Research*, pages 530–544, Yonsei University, Seoul, Republic of Korea, 15–17 Nov 2017. PMLR. URL `https://proceedings.mlr.press/v77/moniz17a.html`.

[187] Ivan Montero, Nikolaos Pappas, and Noah A. Smith. Sentence bottleneck autoencoders from transformer language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1822–1831, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.137. URL `https://aclanthology.org/2021.emnlp-main.137`.

[188] Milton Llera Montero, Casimir J. H. Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. The role of disentanglement in generalisation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=qbH974jKUVy`.

[189] Roser Morante, Chantal van Son, Isa Maks, and Piek Vossen. Annotating perspectives on vaccination. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4964–4973, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.611`.

[190] Aaron Mueller and Mark Dredze. Fine-tuning encoders for improved monolingual and zero-shot polylingual neural topic modeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, pages 3054–3068, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.243. URL `https://aclanthology.org/2021.naacl-main.243`.

[191] Eric Nalisnick, Lars Hertel, and Padhraic Smyth. Approximate inference for deep latent gaussian mixtures. In *NIPS Workshop on Bayesian Deep Learning*, volume 2, page 131, 2016. URL `http://bayesiandeeplearning.org/2016/papers/BDL_20.pdf`.

[192] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL `https://aclanthology.org/D18-1206`.

[193] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2012.07.001. URL `https://www.sciencedirect.com/science/article/pii/S0004370212000793`.

[194] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1113. URL `https://aclanthology.org/D14-1113`.

[195] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. Text and code embeddings by contrastive pre-training. *OpenAI Blog*, 2022. URL `https://openai.com/blog/introducing-text-and-code-embeddings`.

[196] Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. Improving topic models with latent feature word representations. *Transactions of the*

Association for Computational Linguistics, 3:299–313, 2015. doi: 10.1162/tacl_a_00140. URL https://aclanthology.org/Q15-1022.

[197] Tan Nguyen, Vai Suliafu, Stanley Osher, Long Chen, and Bao Wang. Fmmformer: Efficient and flexible transformer via decomposed near-field and far-field attention. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29449–29463. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/f621585df244e9596dc70a39b579efb1-Paper.pdf.

[198] OpenAI. Gpt-4 technical report. *Technical report, OpenAI*, 2023. URL https://arxiv.org/abs/2303.08774.

[199] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022. URL https://openreview.net/forum?id=TG8KACxEON.

[200] Maria Pacheco, Tunazzina Islam, Monal Mahajan, Andrey Shor, Ming Yin, Lyle Ungar, and Dan Goldwasser. A holistic framework for analyzing the COVID-19 vaccine debate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5821–5839, Seattle, United States, July 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.naacl-main.427.

[201] John Paisley, David Blei, and Michael Jordan. Variational bayesian inference with stochastic search. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 1367–1374, New York, NY, USA, July 2012. Omnipress. ISBN 978-1-4503-1285-1. URL https://icml.cc/2012/papers/687.pdf.

[202] John Paisley, Chong Wang, David M. Blei, and Michael I. Jordan. Nested hierarchical dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):256–270, 2015. doi: 10.1109/TPAMI.

2014.2318728. URL https://ieeexplore.ieee.org/document/6802355?arnumber=6802355.

[203] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008. ISSN 1554-0669. doi: 10.1561/1500000011. URL http://dx.doi.org/10.1561/1500000011.

[204] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7198–7211. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/50905d7b2216bfeccb5b41016357176b-Paper.pdf.

[205] Yookoon Park, Jaemin Cho, and Gunhee Kim. A hierarchical latent structure for variational conversation modeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1792–1801, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1162. URL https://aclanthology.org/N18-1162.

[206] Elise Paul, Andrew Steptoe, and Daisy Fancourt. Attitudes towards vaccines and intention to vaccinate against covid-19: Implications for public health communications. *The Lancet Regional Health - Europe*, 1:100012, 2021. ISSN 2666-7762. doi: https://doi.org/10.1016/j.lanepe.2020.100012. URL https://www.sciencedirect.com/science/article/pii/S2666776220300120.

[207] Nicole Peinelt, Dong Nguyen, and Maria Liakata. tBERT: Topic models and BERT joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.630. URL https://aclanthology.org/2020.acl-main.630.

[208] Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (05):8600–8607, Apr. 2020. doi: 10.1609/aaai.v34i05.6383. URL https://ojs.aaai.org/index.php/AAAI/article/view/6383.

[209] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL `https://aclanthology.org/D14-1162`.

[210] Gabriele Pergola, Lin Gui, and Yulan He. Tdam: A topic-dependent attention model for sentiment analysis. *Information Processing & Management*, 56(6):102084, 2019. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2019.102084. URL `https://www.sciencedirect.com/science/article/pii/S0306457319305461`.

[211] Gabriele Pergola, Elena Kochkina, Lin Gui, Maria Liakata, and Yulan He. Boosting low-resource biomedical QA via entity-aware masking strategies. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1977–1985, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.169. URL `https://aclanthology.org/2021.eacl-main.169`.

[212] Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1161. URL `https://aclanthology.org/P17-1161`.

[213] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL `https://aclanthology.org/N18-1202`.

[214] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong

Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250.

[215] Stanislav Pidhorskyi, Donald A. Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Pidhorskyi_Adversarial_Latent_Autoencoders_CVPR_2020_paper.html.

[216] Mohammad Taher Pilehvar and Nigel Collier. De-conflated semantic representations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1680–1690, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1174. URL https://aclanthology.org/D16-1174.

[217] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050.

[218] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *Technical report, OpenAI*, 2018. URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

[219] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[220] Colin Raffel and Daniel P. W. Ellis. Feed-forward networks with attention can solve some long-term memory problems. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Workshop Track Proceedings*, 2016. URL https://arxiv.org/pdf/1512.08756.pdf.

[221] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[222] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12630–12641. IEEE Computer Society, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/papers/Raistrick_Infinite_Photorealistic_Worlds_Using_Procedural_Generation_CVPR_2023_paper.pdf.

[223] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

[224] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David P. Kreil, Michael K. Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

[225] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

[226] Mehdi Rezaee and Francis Ferraro. A discrete variational recurrent topic model without the reparametrization trick. In H. Larochelle, M. Ranzato,

R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13831–13843. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/9f1d5659d5880fb427f6e04ae500fc25-Paper.pdf.

[227] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/rezende15.html.

[228] Oliver Richter and Roger Wattenhofer. Normalized attention without probability cage. *arXiv preprint arXiv:2005.09561*, 2020. URL https://arxiv.org/abs/2005.09561.

[229] Miguel Rios, Wilker Aziz, and Khalil Sima'an. Deep generative model for joint alignment and word representation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1011–1023, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1092. URL https://aclanthology.org/N18-1092.

[230] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. Reasoning about entailment with neural attention. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1509.06664.

[231] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3317-7. doi: 10.1145/2684822.2685324. URL http://doi.acm.org/10.1145/2684822.2685324.

[232] Sara Rosenthal, Noura Farra, and Preslav Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2088. URL https://aclanthology.org/S17-2088.

[233] Frank Rosner, Alexander Hinneburg, Michael Röder, Martin Nettling, and Andreas Both. Evaluating topic coherence measures. *CoRR*, abs/1403.6397, 2014. URL `https://arxiv.org/abs/1403.6397`.

[234] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533, 1986. URL `https://www.nature.com/articles/323533a0#citeas`.

[235] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1044. URL `https://aclanthology.org/D15-1044`.

[236] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=9Vrb9D0WI4`.

[237] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035, 2019. URL `https://cdn.aaai.org/ojs/4160/4160-13-7214-1-10-20190705.pdf`.

[238] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954. URL `https://ieeexplore.ieee.org/document/9363924`.

[239] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, pages 7881–7892. Association for Computational Linguistics, July 2020. doi: 10.18653/v1/2020.acl-main.704. URL `https://aclanthology.org/2020.acl-main.704`.

[240] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=HJ0UKP9ge`.

[241] Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017. doi: 10.1609/aaai.v31i1.10983. URL `https://ojs.aaai.org/index.php/AAAI/article/view/10983`.

[242] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann Lecun. Pedestrian detection with unsupervised multi-stage feature learning. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, page 3626–3633, USA, 2013. IEEE Computer Society. ISBN 9780769549897. doi: 10.1109/CVPR.2013.465. URL `https://doi.org/10.1109/CVPR.2013.465`.

[243] Karin Sevegnani, David M. Howcroft, Ioannis Konstas, and Verena Rieser. OTTers: One-turn topic transitions for open-domain dialogue. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2492–2504, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.194. URL `https://aclanthology.org/2021.acl-long.194`.

[244] Uri Shaham, Kelly P. Stanton, Henry Li, Ronen Basri, Boaz Nadler, and Yuval Kluger. Spectralnet: Spectral clustering using deep neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL `https://openreview.net/forum?id=HJ_aoCyRZ`.

[245] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 464–468, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2074. URL `https://aclanthology.org/N18-2074`.

[246] Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. TaxoClass: Hierarchical multi-label text classification using only class names. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4239–4249, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.335. URL `https://aclanthology.org/2021.naacl-main.335`.

[247] Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. Improving variational encoder-decoders in dialogue generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11960. URL `https://ojs.aaai.org/index.php/AAAI/article/view/11960`.

[248] Bei Shi, Wai Lam, Shoaib Jameel, Steven Schockaert, and Kwun Ping Lai. Jointly learning word embeddings and latent topics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 375–384, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350228. doi: 10.1145/3077136.3080806. URL `https://doi.org/10.1145/3077136.3080806`.

[249] Prasha Shrestha, Suraj Maharjan, Dustin Arendt, and Svitlana Volkova. Learning from dynamic user interaction graphs to forecast diverse social behavior. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2033–2042, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369763. doi: 10.1145/3357384.3358043. URL `https://doi.org/10.1145/3357384.3358043`.

[250] Junaid Shuja, Eisa Alanazi, Waleed Alasmary, and Abdulaziz Alashaikh. Covid-19 open source data sets: A comprehensive survey. *Applied Intelligence*, 51(3):1296–1325, mar 2021. ISSN 0924-669X. doi: 10.1007/s10489-020-01862-6. URL `https://doi.org/10.1007/s10489-020-01862-6`.

[251] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Ku-

maran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. doi: 10.1126/science.aar6404. URL https://www.science.org/doi/abs/10.1126/science.aar6404.

[252] Prateek Sircar, Aniket Chakrabarti, Deepak Gupta, and Anirban Majumdar. Distantly supervised aspect clustering and naming for E-commerce reviews. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 94–102, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-industry.12. URL https://aclanthology.org/2022.naacl-industry.12.

[253] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170.

[254] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf.

[255] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf.

[256] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017. doi: 10.1609/aaai.v31i1.11164. URL https://ojs.aaai.org/index.php/AAAI/article/view/11164.

[257] P. K. Srijith, Michal Lukasik, Kalina Bontcheva, and Trevor Cohn. Longitudinal modeling of social media with hawkes process based on users and networks. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, page 195–202, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349932. doi: 10.1145/3110025.3110107. URL `https://doi.org/10.1145/3110025.3110107`.

[258] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=BybtVK9lg`.

[259] Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL `https://proceedings.neurips.cc/paper/2015/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf`.

[260] Bin Sun, Shaoxiong Feng, Yiwei Li, Jiamou Liu, and Kan Li. Generating relevant and coherent dialogue responses using self-separated conditional variational AutoEncoders. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5624–5637, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.437. URL `https://aclanthology.org/2021.acl-long.437`.

[261] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer, 2019. URL `https://link.springer.com/chapter/10.1007/978-3-030-32381-3_16`.

[262] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL `https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf`.

[263] E. G. Tabak and Cristina V. Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013. doi: https://doi.org/10.1002/cpa.21423. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21423`.

[264] Esteban G. Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8 (1):217–233, 2010. ISSN 1539-6746. doi: 10.4310/CMS.2010.v8.n1.a11. URL `https://content.intlpress.com/journal/CMS/article/11496/info`.

[265] Xingwei Tan, Gabriele Pergola, and Yulan He. Event temporal relation extraction with Bayesian translational model. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1125–1138, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL `https://aclanthology.org/2023.eacl-main.80`.

[266] Yaling Tao, Kentaro Takagi, and Kouta Nakata. Clustering-friendly representation learning via instance discrimination and feature decorrelation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=e12NDM7wkEY`.

[267] Yee Teh, Michael Jordan, Matthew Beal, and David Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL `https://proceedings.neurips.cc/paper/2004/file/fb4ab556bc42d6f0ee0f9e24ec4d1af0-Paper.pdf`.

[268] Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.28. URL `https://aclanthology.org/2021.naacl-main.28`.

[269] Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. Word meaning in context: A simple and effective vector model. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1134–1143, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing. URL `https://aclanthology.org/I11-1127`.

[270] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume. Lample. Llama: Open and efficient foundation language models, 2023. URL `https://research.facebook.com/publications/llama-open-and-efficient-foundation-language-models/`.

[271] Ke Tran, Arianna Bisazza, and Christof Monz. Recurrent memory networks for language modeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 321–331, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1036. URL `https://aclanthology.org/N16-1036`.

[272] Nam N Tran and Joon Lee. Online reviews as health data: Examining the association between availability of health care services and patient star ratings exemplified by the yelp academic dataset. *JMIR Public Health Surveill*, 3 (3):e43, Jul 2017. ISSN 2369-2960. doi: 10.2196/publichealth.7001. URL `http://publichealth.jmir.org/2017/3/e43/`.

[273] Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10401–10412. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/trauble21a.html`.

[274] Alan Mathison Turing. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, 2 (1):230–230, 1937. URL `https://londmathsoc.onlinelibrary.wiley.com/doi/abs/10.1112/plms/s2-42.1.230`.

[275] Keyon Vafa, Yuntian Deng, David Blei, and Alexander Rush. Rationales for sequential predictions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10314–10332, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.807. URL `https://aclanthology.org/2021.emnlp-main.807`.

[276] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[277] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[278] Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/37bc2f75bf1bcfe8450a1a41c200364c-Paper.pdf.

[279] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

[280] Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6623.

[281] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/29921001f2f04bd3baee84a12e98098f-Paper.pdf.

[282] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1–2):1–305, jan 2008. ISSN 1935-8237. doi: 10.1561/2200000001. URL https://doi.org/10.1561/2200000001.

[283] Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z. Pan. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9122–9129, Apr. 2020. doi: 10.1609/aaai.v34i05.6447. URL `https://ojs.aaai.org/index.php/AAAI/article/view/6447`.

[284] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL `https://aclanthology.org/W18-5446`.

[285] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf`.

[286] Jingjing Wang, Jie Li, Shoushan Li, Yangyang Kang, Min Zhang, Luo Si, and Guodong Zhou. Aspect sentiment classification with both word-level and clause-level attention networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4439–4445. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/617. URL `https://doi.org/10.24963/ijcai.2018/617`.

[287] Kexin Wang, Nils Reimers, and Iryna Gurevych. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.59. URL `https://aclanthology.org/2021.findings-emnlp.59`.

[288] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by

weakly-supervised contrastive pre-training. In *Microsoft Blog*, December 2022. URL https://www.microsoft.com/en-us/research/publication/text-embeddings-by-weakly-supervised-contrastive-pre-training/.

[289] Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Ranran Haoran Zhang, Weili Liu, Aabhas Chauhan, Yingjun Guan, Bangzheng Li, Ruisong Li, Xiangchen Song, Yi Fung, Heng Ji, Jiawei Han, Shih-Fu Chang, James Pustejovsky, Jasmine Rah, David Liem, Ahmed ELsayed, Martha Palmer, Clare Voss, Cynthia Schneider, and Boyan Onyshkevych. COVID-19 literature knowledge graph construction and drug repurposing report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 66–77, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-demos.8. URL https://aclanthology.org/2021.naacl-demos.8.

[290] Xinyu Wang, Lin Gui, and Yulan He. Document-level multi-event extraction with event proxy nodes and hausdorff distance minimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10118–10133, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.563.

[291] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

[292] Anuradha Welivita, Yubo Xie, and Pearl Pu. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.96. URL https://aclanthology.org/2021.emnlp-main.96.

[293] Sean Welleck, Kianté Brantley, Hal Daumé Iii, and Kyunghyun Cho. Non-monotonic sequential text generation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*,

pages 6716–6726. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/welleck19a.html`.

[294] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL `https://aclanthology.org/2020.emnlp-demos.6`.

[295] Chuhan Wu, Fangzhao Wu, Sixing Wu, Junxin Liu, Zhigang Yuan, and Yongfeng Huang. THU_NGN at SemEval-2018 task 3: Tweet irony detection with densely connected LSTM and multi-task learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 51–56, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1006. URL `https://aclanthology.org/S18-1006`.

[296] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Hi-transformer: Hierarchical interactive transformer for efficient and effective long document modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 848–853, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.107. URL `https://aclanthology.org/2021.acl-short.107`.

[297] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. URL `https://arxiv.org/abs/1609.08144`.

[298] Shuai Xiao, Junchi Yan, Xiaokang Yang, Hongyuan Zha, and Stephen Chu. Modeling the intensity function of point process via recurrent neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb.

2017. doi: 10.1609/aaai.v31i1.10724. URL https://ojs.aaai.org/index.php/AAAI/article/view/10724.

[299] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 478–487, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/xieb16.html.

[300] Hongteng Xu, Wenlin Wang, Wei Liu, and Lawrence Carin. Distilled wasserstein learning for word embedding and topic modeling. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1722–1731. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7443-distilled-wasserstein-learning-for-word-embedding-and-topic-modeling.pdf.

[301] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/xuc15.html.

[302] Lu Xu, Hao Li, Wei Lu, and Lidong Bing. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.183. URL https://aclanthology.org/2020.emnlp-main.183.

[303] Hanqi Yan, Lin Gui, and Yulan He. Hierarchical Interpretation of Neural Text Classification. *Computational Linguistics*, 48(4):987–1020, 12 2022. ISSN 0891-2017. doi: 10.1162/coli_a_00459. URL https://doi.org/10.1162/coli_a_00459.

[304] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining

for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

[305] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1174. URL https://aclanthology.org/N16-1174.

[306] Ziyi Yang, Yinfei Yang, Daniel Cer, Jax Law, and Eric Darve. Universal sentence representation learning with conditional masked language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6216–6228, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.502. URL https://aclanthology.org/2021.emnlp-main.502.

[307] Deniz Yuret. KU: Word sense disambiguation by substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 207–214, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/S07-1044.

[308] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf.

[309] Sayyed Zahiri and Jinho D. Choi. Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks. In *Proceedings of the AAAI-18 Workshop on Affective Content Analysis*, AFFCON'18, pages 44–51, New Orleans, LA, 2018. URL https://sites.google.com/view/affcon18.

[310] Min Zeng, Yisen Wang, and Yuan Luo. Dirichlet latent variable hierarchical recurrent encoder-decoder in dialogue generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1267–1272, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1124. URL `https://aclanthology.org/D19-1124`.

[311] Chen Zhang, Qiuchi Li, and Dawei Song. Aspect-based sentiment classification with aspect-specific graph convolutional networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4568–4578, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1464. URL `https://aclanthology.org/D19-1464`.

[312] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=Sy8gdB9xx`.

[313] Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. Supporting clustering with contrastive learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5419–5430. Association for Computational Linguistics, June 2021. doi: 10.18653/v1/2021.naacl-main.427. URL `https://aclanthology.org/2021.naacl-main.427`.

[314] Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5415–5421. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/752. URL `https://doi.org/10.24963/ijcai.2019/752`.

[315] Haidong Zhang, Wancheng Ni, Meijing Zhao, and Ziqi Lin. Cluster-gated convolutional neural network for short text classification. In *Pro-*

ceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 1002–1011, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1094. URL https://aclanthology.org/K19-1094.

[316] Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. Discovering new intents with deep aligned clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14365–14373, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17689.

[317] Meishan Zhang, Yue Zhang, and Duy-Tin Vo. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 612–621, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1073. URL https://aclanthology.org/D15-1073.

[318] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

[319] W. Zhang, C. Dong, J. Yin, and J. Wang. Attentive representation learning with adversarial training for short text clustering. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5196–5210, nov 2022. ISSN 1558-2191. doi: 10.1109/TKDE.2021.3052244. URL https://www.computer.org/csdl/journal/tk/2022/11/09328333/1qutQsyD04w.

[320] Wenjia Zhang, Lin Gui, Rob Procter, and Yulan He. Newsquote: A dataset built on quote extraction and attribution for expert recommendation in fact-checking. *International AAAI Conference on Web and Social Media (ICWSM)*, 2023. URL https://workshop-proceedings.icwsm.org/abstract.php?id=2023_22.

[321] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

[322] Xiongyi Zhang, Jan-Willem van de Meent, and Byron Wallace. Disentangling representations of text by masking transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 778–791, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.60. URL `https://aclanthology.org/2021.emnlp-main.60`.

[323] Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul N. Bennett, and Saurabh Tiwary. Transformer-xh: Multi-evidence reasoning with extra hop attention. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=r1eIiCNYwS`.

[324] He Zhao, Lan Du, Wray Buntine, and Mingyuan Zhou. Inter and intra topic structure learning with word embeddings. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5892–5901. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/zhao18a.html`.

[325] He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. SpanMlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3239–3248, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.296. URL `https://aclanthology.org/2020.acl-main.296`.

[326] Runcong Zhao, Lin Gui, and Yulan He. Cone: Unsupervised contrastive opinion extraction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1066–1075, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591650. URL `https://doi.org/10.1145/3539618.3591650`.

[327] Runcong Zhao, Lin Gui, Hanqi Yan, and Yulan He. Tracking brand-associated polarity-bearing topics in user reviews. *Transactions of the Association for Computational Linguistics*, 11:404–418, 2023. doi: 10.1162/tacl_a_00555. URL `https://aclanthology.org/2023.tacl-1.24`.

[328] Peixiang Zhong, Di Wang, and Chunyan Miao. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of*

the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 165–176, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1016. URL https://aclanthology.org/D19-1016.

[329] Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu. Deep recurrent models with fast-forward connections for neural machine translation. *Transactions of the Association for Computational Linguistics*, 4:371–383, 2016. doi: 10.1162/tacl_a_00105. URL https://aclanthology.org/Q16-1027.

[330] Naitian Zhou and David Jurgens. Condolence and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 609–626, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.45. URL https://aclanthology.org/2020.emnlp-main.45.

[331] Lixing Zhu, Yulan He, and Deyu Zhou. Neural opinion dynamics model for the prediction of user-level stance dynamics. *Information Processing & Management*, 57(2):102031, 2020. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2019.03.010. URL https://www.sciencedirect.com/science/article/pii/S0306457318308604.

[332] Lixing Zhu, Yulan He, and Deyu Zhou. A neural generative model for joint learning topics and topic-specific word embeddings. *Transactions of the Association for Computational Linguistics*, 8:471–485, 2020. doi: 10.1162/tacl_a_00326. URL https://aclanthology.org/2020.tacl-1.31.

[333] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.125. URL https://aclanthology.org/2021.acl-long.125.

[334] Lixing Zhu, Zheng Fang, Gabriele Pergola, Robert Procter, and Yulan He. Disentangled learning of stance and aspect topics for vaccine attitude detection in social media. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, pages 1566–1580, Seattle, United States, July 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.naacl-main.112`.

[335] Lixing Zhu, Runcong Zhao, Gabriele Pergola, and Yulan He. Disentangling aspect and stance via a Siamese autoencoder for aspect clustering of vaccination opinions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1827–1842, Toronto, Canada, July 2023. Association for Computational Linguistics. URL `https://aclanthology.org/2023.findings-acl.115`.

[336] Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2438–2448, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL `https://aclanthology.org/C16-1230`.

[337] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2), feb 2018. ISSN 0360-0300. doi: 10.1145/3161603. URL `https://doi.org/10.1145/3161603`.