**Noname manuscript No.**
(will be inserted by the editor)

# TweetLID: A Benchmark for Tweet Language Identification

**Arkaitz Zubiaga**[1] · **Iñaki San Vicente**[2] · **Pablo Gamallo**[3] · **José Ramom Pichel**[4] · **Iñaki Alegria**[5] · **Nora Aranberri**[5] · **Aitzol Ezeiza**[5] · **Víctor Fresno**[6]

Received: date / Accepted: date

**Abstract** Language identification, as the task of determining the language a given text is written in, has progressed substantially in recent decades. However, three main issues remain still unresolved: (i) distinction of similar languages, (ii) detection of multilingualism in a single document, and (iii) identifying the language of short texts. In this paper, we describe our work on the development of a benchmark to encourage further research in these three directions, set forth an evaluation framework suitable for the task, and make a dataset of annotated tweets publicly available for research purposes. We also describe the shared task we organized to validate and assess the evaluation framework and dataset with systems submitted by seven different participants, and analyze the performance of these systems. The evaluation of the results submitted by the participants of the shared task helped us shed some light on the shortcomings of state-of-the-art language identification systems, and gives insight into the extent to which the brevity, multilingualism, and language similarity found in texts exacerbate the performance of language identifiers. Our dataset with nearly 35,000 tweets and the evaluation framework provide researchers and practitioners with suitable resources to further study the aforementioned issues on language identification within a common setting that enables to compare results with one another.

**Keywords** language identification · tweets · short texts · multilingualism · similar languages

## 1 Introduction

Recent research shows that while Twitter's predominant language was English in its early days, the global growth and adoption of the social media platform in recent years has increased the diversity in the use of languages [36]. This has in turn fostered an increasing interest of the scientific community in automatically guessing

[1] University of Warwick, [2] Elhuyar, [3] USC
[4] imaxin|software, [5] University of the Basque Country, [6] UNED
tweetlid@elhuyar.com

the languages of tweets [10]. The identification of the language of a tweet is crucial for the subsequent application of widely used NLP tools such as machine translation [27], sentiment analysis [1,34], Named Entity Recognition (NER) [37], entity linking [22,11], text summarization [54,78], and lexical [2] and syntactic normalization [29], among others. The main problem lies in that this kind of NLP tools tend to be crafted with resources specifically trained for a language or some languages. Hence, these tools cannot deal with unknown languages unless suitable resources are developed. This makes language identification a crucial task especially in multilingual environments such as Twitter, where accurately identifying the language of a tweet enables the application of NLP resources suitable to the language in question.

Twitter itself does provide a language id along with each tweet's metadata, but as we show in this article it leaves much to be desired in terms of accuracy. Besides, it is intended to detect major languages, and does not identify other languages with lesser presence on the platform such as Catalan, Basque or Galician, which account for millions of native speakers within the Iberian Peninsula. In this work, we set out to study the development of language identification systems that deal with more complex situations, including the aforementioned shortcomings of Twitter. To that end, we first review the related work on language identification and the issues that remain unresolved as of today. Then, we introduce a benchmark dataset and evaluation framework that enables to evaluate different language identification systems, dealing with three of the most important issues that are not resolved: (i) distinction of similar languages, (ii) detection of multilingualism in a single document, and (iii) identifying the language of short texts.

To develop and validate such a benchmark dataset and evaluation framework, we have organized a shared task on tweet language identification (TweetLID), and invited researchers to submit their language identification systems. The task focused on the five most spoken languages of the Iberian Peninsula (Spanish, Portuguese, Catalan, Basque and Galician), and English. These languages are likely to co-occur along with many news and events relevant to the Iberian Peninsula, and thus an accurate identification of the language is key to make sure that we use the appropriate resources for the linguistic processing. This task has intended to bring together contributions from researchers and practitioners in the field, to develop and compare tweet language identification systems designed for the aforementioned languages, which can potentially later be extended to a wider variety of languages. The task meets the aforementioned unresolved issues, given that (i) the task includes four Romance languages which are somewhat similar to one another, (ii) tweets can often be multilingual, and (iii) tweets are short by nature.

This research aims to satisfy the lack of both a benchmark dataset and an evaluation framework to compare different language identification systems. This dataset can be further used by interested researchers and practitioners to make progress in the development of tweet language identification systems.

In this paper, we introduce the benchmark dataset and evaluation framework that enabled the organization of the shared task, which is also made publicly available for research purposes. Then, we analyze and discuss the performance of the different participants of the shared task, which brings to light the most challenging aspects encountered by the participants and need to be addressed in future work. We end

by discussing the main objectives that language identification for short texts should pursue in the next years.

This paper substantially extends the overview article we published with the proceedings of the TweetLID workshop [77]. In this extended paper, we provide an extensive review of the literature, and perform a detailed analysis of the results, by looking among others at numerous aspects relevant to the task, including the three unresolved issues, namely the brevity of texts, multilingualism, and similar languages. Moreover, this paper discusses the achievements and limitations of the presented systems, summarizing the challenges that are still open for future work.

## 2 Language Identification

Language identification consists in determining the language a text is written in. It has usually been tackled as a classification problem in previous research, often assuming that a document is entirely written in a single language. The best known approaches make use of n-grams to learn the model for each of the languages, as well as to represent each of the documents to be categorized into one of the languages [12]. A language identification system is usually defined as a text classification task [61].

Here we focus on language identification for short texts, more specifically tweets, which is still in its infancy as a research field. Tweets present different characteristics that make the language identification task more challenging. These include that:

- The brevity of the tweets implies that there is very little content that helps to determine the language being used.
- The system allows to use different features along with the content, which do not usually reflect the language of the text. These features include user mentions, hashtags, or retweets, among others.
- Users tend to shorten and/or encode many words in the form of chatspeak, while also introducing typos and misspellings, which deviates the text from its standard spelling.

Provided the aforementioned characteristics inherent in tweets, the language identification for these short texts involves a number of extra challenges that were not considered in other language identification tasks for standard documents such as news stories, books, or even the Web.

## 3 Related Work

In this section, we review previous work in the literature. We start with the historical background of the research in the field of language identification. Then, we summarize the findings of several comparative studies, and continue by discussing the different directions that research in this field has taken, including language identification for web pages, word level language identification, and language identification for short texts and tweets. We then discuss recent shared tasks that were related to the objectives of TweetLID, and conclude the section by enumerating and discussing the state-of-the-art of the main challenges that our work deals with.

### 3.1 Historical Background

Language identification has attracted a substantial interest in the scientific community in recent decades. While the task was first studied within the community of translators [5,51,30,26] mostly in the 1980s, it started to be more widely studied within the machine learning and natural language processing communities in the 1990s [12,16].

Early work on language identification from texts relied on manually defining rules that could be useful in the development of computational tools. For instance, Beesley [5] proposed relying on language-specific characters to distinguish certain languages, such as ñ or ü for Spanish, or ã for Portuguese. Beesley suggested that such an approach could perform reasonably well for certain languages. However, this approach could perform well for reasonably long and correctly spelled texts in a small set of languages, but more sophisticated techniques might be needed in other scenarios. Later, Cavnar and Trenkle [12] introduced one of the earliest and most frequently used approaches to language identification in texts: TextCat. Their system computes the n-grams from an input text, and compares the n-grams to the models learned for each of the target languages. The system computes the distance measures with respect to each target language, to assign the language with the lowest distance. This approach achieved 99.8% correct classification rate on Usenet newsgroup articles. Dunning [16] developed a language identification system using Markov models and a Bayesian classifier. The classifier looks for sequences of characters and words that are unique for each language in the training set, to find similar patterns in the test set. He showed that with only 50k characters of training data, the system could achieve up to 92% accuracy values when identifying the language for short texts of 20 characters. The accuracy increased to more than 99% with larger training sets and test strings with more than 100 characters. He pointed out five key conditions that determine the performance of a language identification system: (i) how the test strings are picked, (ii) the amount of training material available, (iii) the size of the strings to be identified, (iv) the number of languages to be identified, and (v) whether there is a correlation between domain and language.

In another early attempt, Prager [59] introduced Linguini, a language identification system which uses n-grams and words as features. The system achieved high performance for classification of monolingual documents in 20 different languages, but its performance dropped significantly for short texts. The author also discussed the applicability of the method to bilingual and trilingual documents. Among the different features studied, 4-grams showed to be the best length for n-grams, and words of unrestricted length did better than considering only short words. The combination of both, 4-grams and words of unrestricted length, performed best. More recently, Lui and Baldwin [40] developed a method suited to cross-domain language identification. It relies on information gain to identify the features that are strongly predictive of language across domains. Building a feature set from 50,000 documents in 97 languages across 5 datasets, the authors showed that the proposed method can outperform well-known systems such as TextCat [12] when applied to different domains. Finally, Lui and Baldwin [41] released langid.py, an off-the-shelf language identification script developed in Python. The script is developed using a Naive Bayes algorithm that re-

lies on n-grams extracted from texts to identify the language, and is intended to be easy-to-use and applicable to different domains.

### 3.2 Comparison Studies

As research in language identification systems made progress, some researchers also conducted comparison studies to find the approaches that work best. Grefenstette [21] compared two language identification approaches. One using character trigrams as features, and the other one using common short words as features. Their experiments on corpora in 10 European languages showed that either of the compared approaches achieves high accuracy for long texts with more than 50 words, but that trigrams are much more robust for shorter texts. Padró and Padró [56] compared three statistical methods for language identification: Markov Models, Trigram Frequency Vectors, and n-gram text categorization. They used corpora in 6 different languages for their experiments. They found that Markov Models performed best among the three approaches under study. While the size of the training set did not have a huge impact in the system performance when the training set had at least 50,000 words, they found significant differences in performance when the texts to be classified were very short. Baldwin and Lui [3] describe a set of experiments comparing different language identification techniques on three web document datasets. Comparing 1-Nearest Neighbors (1-NN), Naive Bayes, and Support Vector Machines (SVM) with different similarity measures. They found that the most consistent model overall is either a simple 1-NN model with cosine similarity, or an SVM with a linear kernel, using a byte bigram or trigram document representation. They posit that the task becomes increasingly challenging as the number of target languages increases, the size of the training data decreases, and the length of the documents is shorter.

### 3.3 Web-Based Approaches

The emergence of the Web, as an information source that gathers a myriad of documents in an endless number of languages, attracted also a community of researchers to studying language identification approaches in this scenario. Kikui [31] described a language identification system for online documents. The system was implemented using language models, and could deal with 9 language and 11 coding systems from Eastern Asia and Western Europe. Their experiments on 640 online documents led to a level of accuracy over 95%. On another study on language identification for web pages, Martins and Silva [45] used the system implemented by [12], complemented with heuristics that specifically deal with HTML markup, and a new similarity measure. They used the web page language identification system to build a search engine that only indexes web content in Portuguese. Their system achieved 99% accuracy in distinguishing Portuguese from the rest of the languages. Xafopoulos et al. [73] used Hidden Markov Models (HMM) to model character sequences in web documents. Their experiments with web documents in 5 European languages, achieving accuracy values of up to 97%. Baykan et al. [4] studied the feasibility of determining the language of the content of a web page by only looking at its URL, i.e., without having

to download its content. They built a classifier based on keywords extracted from URLs, which was tested on a collection of web pages in 5 languages, achieving 90% in terms of F1 measure. Xia et al. [74] study the suitability of existing language identification techniques to collections including documents written in one of hundreds of languages, which they motivate as being closer to the nature of the Web. Using the ODIN database[1] for the experiments, which includes documents in nearly a thousand languages, they found that well-known language identification techniques achieved performance values as low as 55%. They introduced a new method which uses context within the document, and formulated the task as a coreference resolution problem, achieving higher performance than using existing techniques for collections with a large number of languages and small training data. Similar to ODIN, the work by Ralf Brown [7,8] has focused on expanding the number of languages considered simultaneously (developing a language identification system for over 1,100 languages). Alongside these works, the Crubadan Project, led by Kevin Scannell [60], aimed at building a large corpus for under-resourced languages using the Web as a source. The project led to the creation of a corpus in more than 400 languages, especially intended for the development of linguistic resources for under-resourced languages.

### 3.4 Word Level Strategies

Motivated by the fact that there are many multilingual speakers who often switch between languages within a sentence, in recent years there is also an increasing interest in the study of word level language identification, i.e., determining what language each word of a sentence is written in. Nguyen and Doğruöz [52] built a dataset from a Turkish-Dutch community of users, where users mix these two languages, occasionally mixing it with English too. By annotating the language of single words, they experimented with Conditional Random Fields (CRF), which they proved effective at nearly 98% accuracy when using the previous and next tokens to add context to each word. Gella et al. [18] studied word level language identification for 28 languages, where the system does not know a priori which two languages might co-occur in a text. They defined different heuristics, applied to existing language identification tools such as langid.py and linguini. The heuristics include, for instance, assuming that code-mixing is only likely to occur between certain pairs of languages, but not any possible pair. Their system outperformed existing language identification techniques which are not designed to deal with code-mixed texts, but tends to confuse between languages which are linguistically related. King and Abney [32] described a weakly supervised language identification system which can be trained using monolingual text samples. Using n-grams as the features to represent the texts, they showed that Conditional Random Fields (CRF) with Generalized Expectation (GE) [15] criteria performed best. The major issue they encountered in the word level identification task were the Named Entities (NE) mentioned in the text, which are very difficult to identify when the language is unknown a priori. They conclude suggesting that a word level language identification system could be built in two steps, the first step

[1] http://odin.linguistlist.org/

being the high level identification of languages used in a text, and the second step being the specific assignment of language labels to words.

### 3.5 Tweets/Short Messages

Little work has been done on language identification of short texts. Research in this direction has increased especially in recent years, with the advent of social media and microblogs. Tromp and Pechenizkiy [68] proposed a graph-based n-gram approach for tweet language identification. Using Twitter datasets with monolingual tweets in six languages, they achieved performances between 95% and 98%. Vogel and Tresner-Kirsch [71] extend the work by Tromp and Pechenizkiy by proposing several linguistically-motivated modifications to their algorithm and achieving 99.8% accuracy.

Laboreiro et al. [35] used a Bayesian classifier to distinguish between European and Brazilian variants of tweets written in Portuguese language, achieving 95% accuracy. Winkelmolen and Mascardi [72] also describe a Bayesian classifier that performs well on very short texts and made experiments on film subtitles in 22 languages. The work by Murthy and Kumar [49] deal with short texts, and are especially interested in satisfying the scarcity of research in language identification for a variety of Indian languages, including Hindi, Bengali, Marathi, Punjabi, Oriya, Telugu, Tamil, Malayalam and Kannada. Bergsma et al. [6] studied different language identification techniques on Twitter datasets with tweets in 9 languages which use Cyrillic, Arabic, and Devanagari scripts. Multilingual tweets were annotated with the predominant language in the tweet, and hence multilingualism was not considered. Given that the dataset includes 3 languages in each of the alphabets, they divide the task into 3 smaller subtasks. They tested three language identification systems, using textual features such as n-grams, and user metadata from Twitter, as well as Wikipedia as an external resource. They showed that by combining n-grams and user metadata, their system can achieve up to 98% accuracy in each subtask that deals with three languages. Goldszmidt et al. [19] tested statistical language identifiers, based on character frequencies, to classify tweets in five different languages by using Wikipedia for training. While they found that Wikipedia is insufficient to represent several idioms used exclusively in social media, they introduced a boot-strapping technique that significantly improves the accuracy of the language identifier. Hammarström [23] described a fine-grained model which stores a large frequency dictionary as well as an affix table and is able to classify with high accuracy short texts of just one word.

Carter et al. [10] investigated language identification on a Twitter dataset with tweets in five major European languages: Dutch, English, French, German, and Spanish. To enrich the textual content of tweets, they use additional context surrounding the tweets: (i) the content of the link being pointed to, (ii) the author of the tweet, (iii) mentions of other users, (iv) context from the tweet that it is replying to, and (v) hashtags. They found the combination of all five features to perform best. In our work, we argue that the collection of such context for each tweet is time-consuming, and makes it impossible to run the language identifier in a timely fashion for a rela-

tively large set of tweets. To account for this, we present a tweet dataset and describe the problem as a task where the language of a tweet has to be determined from its readily available features.

Lui and Baldwin [42] presented an evaluation of several language identification systems applied to tweets. They showed that simple voting over three specific systems consistently outperforms any specific system, and achieves state-of-the-art accuracy on the task. In addition, the authors also defined a semi-automatic method to construct annotated datasets of tweets for evaluating a language identification system.

In a comparative study where a number of well-known language identification systems were tested on a Twitter dataset with tweets in five languages, [14] showed that Cavnar and Trenkle's TextCat [12], retraining its models based on tweets, performed best. This comparison also shows a big difference between training TextCat in tweets (97.4% accuracy), or using its own models (89.5% accuracy). Additionally, [20] and [70] have also studied the application of language identification systems to short texts such as search queries.

### 3.6 Related Shared Tasks

In recent years, there have been several shared tasks on language identification, which are relevant to the shared task we organized at TweetLID. The 2010 Australasian Language Technology (ALTA-2010) organized a workshop and shared task on Multilingual LangID. The dataset for the task was created by Baldwin and Lui [39] from editions of Wikipedia in different languages. In 2013, the workshop on Innovative Use of NLP for Building Educational Applications (BEA8)[2], co-located with NAACL, hosted a shared task on Native Language Identification (NLI). The task consisted in identifying the native language of a writer based solely on a sample of their writing [67]. Another relevant shared task is Language Identification in Code-Switched (CS)[3], which was part of the First Workshop on Computational Approaches to Code Switching, organized within the EMNLP-2014 conference. This shared task focused on short texts having in than one language. Moreover, the shared task Discriminating Similar Languages (DSL-2014)[4], organized within COLING-2014, deals with discriminating between similar languages and language varieties, which is one of the bottlenecks of language identification.

### 3.7 Challenges

Among the little work on the study of language identification techniques for tweets, no research has dealt so far with code-mixing and the identification of multilingualism in tweets, and no special attention has been paid to similar languages in these short texts. Our work looks specifically at these two aspects, multilingualism and similar languages, in the context of short texts.

[2] http://www.cs.rochester.edu/~tetreaul/naacl-bea8.html
[3] http://emnlp2014.org/workshops/CodeSwitch/call.html
[4] http://corporavm.unikoeln.de/vardial/sharedtask.html

Others have looked at additional challenges that can occasionally be also part of a language identification task. Chepovskiy et al. [13] looked at how to deal with language identification of transliterated texts. They explored the ability to identify five Slavic languages from their Latin transliterations. Also, Sibun and Reynar [64] studied the accuracy of language identification systems when applied to scanned images. Our work, instead, assumes that the input sentences are given as texts.

Regarding the identification of similar languages, Ljubešić et al. [38] studied the case of Croatian, which language identification tools find it hard to distinguish from similar languages such as Serbian, Slovenian, or Slovak. By defining a set of rules that specifically characterize the Croatian language, such as identifying the most frequent words, their system outperformed existing tools.

Language identification has progressed significantly in recent years, to the point that the task has been considered solved for certain situations [46], assuming among others that documents are long enough and that are written in a single language. However, the emergence of social media and the chatspeak employed by its users has brought about new previously unseen issues that need to be studied in order to deal with these kinds of texts. Three key issues posited in the literature [63, 24, 69] and that, as of today, cannot be considered solved include: (i) distinguishing similar languages [76], (ii) dealing with multilingual documents [43], and (iii) language identification for short texts [6, 10, 35, 20, 70, 52]. The shared task organized at TweetLID has considered these three unresolved issues, and has enabled participants to compare the performance of their systems in these situations.

## 4 Defining the Tweet Language Identification Task

Within the linguistically diverse nature of social media, and specifically Twitter in our case, we set forth the tweet language identification task as the problem that consists in identifying the language or languages tweets are written in. In this work, we have created a Twitter dataset that enables to study language identifiers in a context where tweets are of multilingual nature, often due to the users' tendency to code-mixing, and there is a high degree of similarity between some of the languages. This dataset has been tested in a shared task, TweetLID [77], which allowed participants to evaluate their language identifiers in a common setting. The dataset and task focused on the most widely used languages of the Iberian Peninsula, which provides an ideal context where news and events are likely to be shared and discussed in multiple languages.

To the end of setting up a common evaluation framework to enable comparison of different language identification systems, we put together an annotated corpus of nearly 35,000 tweets and defined a methodology to evaluate the multi-label output of the language identification systems. Splitting the corpus into a training set with 15k tweets, and a test set with 20k tweets, the participants had a month to develop their language identification systems making use of the training set. They then had 72 hours to work on the test set and submit their results. The shared task consisted of two separate tracks: (1) constrained, where external corpora could not be used for training, and (2) unconstrained, where the use of external corpora was permitted. Each participant could participate with up to two submissions per track.

Besides the challenge of dealing with the short and often informal texts found in tweets, the task considered that a tweet is not necessarily written in a single language. This is especially true in bilingual regions, where speakers that feel equally comfortable with either of their two native languages tend to code-switch between them and mix them in a sentence quite frequently [9,50,57]. Hence, the task also considered a number of cases where the response is not basically one of the languages in the list: (i) a tweet can combine two –or occasionally three– languages in a tweet, e.g., when a tweet has parts in Catalan and Spanish, (ii) given the similarity and cultural proximity between some of the languages, it is not possible to determine which of two –or more– languages a tweet is written in, e.g., some tweets might be written equally in Catalan or Spanish, (iii) despite the geographical restriction of the tweets in the task, it is also likely that tweets in other languages occur, such as French, and (iv) it is not possible to determine which of the 6 languages considered in the task a tweet is written in, e.g., when a tweet only mentions entities, smileys, or onomatopoeias. We will elaborate more on these cases in the next section introducing the dataset and the annotation process.

The dataset includes the five top languages of the Iberian Peninsula, which are spoken in different regions, and four of them –Spanish, Portuguese, Catalan, and Galician– are romance languages originating from Latin and with certain similarities among them, which makes the task more challenging. The fifth language –Basque–, and English, belong to different language families, and therefore are rather different from the rest. Still, their cultural proximity, and the fact that many users in the area are bilingual, entails that they often mix words and spellings across languages. For instance, a Basque native might naturally write something like *"nos vemos, agur!"* (see you later, bye!), when *"nos vemos"* is in Spanish, and *"agur"* is Basque to say good bye; similarly, a Catalan speaker might often misspell the Spanish word *"prueba"* (test) as *"prueva"*, given that the Catalan translation of the word (*"prova"*) is written with *v*. These characteristics are common in bilingual areas, and have been considered in the definition of this task in order to carefully develop the annotation guidelines and to pursue the final annotation of the corpora.

## 5 Creation of a Benchmark Dataset and Evaluation Framework

In this section, we first describe the process we followed to collect data from Twitter, then we explain how we annotated manually the tweets with the language label in question, and finally we describe the evaluation measures we used for the task.

### 5.1 Data Collection

To collect an unrestricted set of tweets, but rather focused on the set of languages within the scope of TweetLID, we relied on geolocation to retrieve tweets posted from areas of interest. We used Twitter's streaming API's `statuses/filter` endpoint to collect geolocated tweets posted within the Iberian Peninsula from March 1 to 31, 2014. While this stream is limited to tweets explicitly providing geolocation

metadata, it allows to track a diverse set of tweets that is not restricted to a specific set of users or domain. Having collected these tweets, we used Nominatim[5] to obtain specific location information for each tweet. Given the coordinates of a tweet as input, Nominatim queries OpenStreetMap for the specific address associated with those coordinates, i.e., region, city, and street (if available) from which the tweet has been sent. This led to the collection of 9.7 million tweets with location details associated. From this set of tweets, we sampled tweets from **Portugal** and the following **3 officially bilingual regions**:

- **Basque Country**, where Basque and Spanish are spoken. Tweets from the province of Gipuzkoa were chosen here to represent the Basque Country.
- **Catalonia**, where Catalan and Spanish are spoken. Tweets from the province of Girona were chosen to represent Catalonia.
- **Galicia**, where Galician and Spanish are spoken. Tweets from the province of Lugo were chosen.

One province was picked from each of the regions to avoid cases such as that of the province of Barcelona in Catalonia, which is much more diverse in terms of languages due to tourism. These three bilingual regions enabled us to sample tweets in Basque, Catalan, Galician, and Spanish, and we could sample Portuguese tweets from Portugal. English is the sixth language in the corpus, which can be found all across the aforementioned regions. For the final corpus to be manually annotated, we picked 10k tweets from each of the bilingual regions, and 5k from Portugal. The tweets picked here had to contain at least one word (i.e., string fully made of a-z characters), so that there is some text, and tweets with e.g. only a link are not considered. The next section describes the manual annotation performed on this corpus with 35k tweets.

### 5.2 Manual Annotation

The collection of 35k tweets resulting from the aforementioned process was then manually annotated. Each of the tweets was associated with its corresponding language code in the manual annotation process. The manual annotation was conducted by annotators who were native or proficient speakers in at least three languages considered in the task. This enabled us to distribute the tweets from each of the four regions to different annotators, so that each annotator was a native or proficient speaker of the languages spoken in the region in question, as well as English.

The annotators were instructed to assign codes to tweets according to the language in which they were written. We asked them to ignore #hashtags and @user mentions, as well as references to NEs in another language. For instance, in the tweet *Acabo de ver el último capítulo de la temporada de 'the walking dead', muy bueno!* (Spanish: I just saw the season finale of 'the walking dead', it's amazing!), only Spanish should be annotated, irrespective of the named entity 'the walking dead' being in English.

[5] http://wiki.openstreetmap.org/wiki/Nominatim

They had to assign codes to the tweets as follows: *eu* for Basque, *ca* for Catalan, *gl* for Galician, *es* for Spanish, *pt* for Portuguese, and *en* for English. When a different language was found in a tweet –e.g., French or German–, they had to annotate it as *other*. Additionally, when the text of a tweet included words that are widely used in any of the languages in the task –e.g., onomatopoeias such as 'jajaja' or 'hahaha', or internationalized words such as 'ok'–, which makes it impossible to determine the language being used in that specific case, they were asked to annotate it as *und*(eterminable). These eight cases —i.e., eu, ca, gl, es, pt, en, other, und– constitute all the options for **monolingual tweets**.

In the above situations, the annotators had to mark a tweet as either being written in one of the 6 languages, *other* or *und*. However, two more cases were identified and included in the annotation guidelines: ambiguous tweets, and multilingual tweets.

**Ambiguous tweets** were defined as those that can be categorized into the list of languages being considered, but may have been written in at least two of them. Given the similarity and cultural proximity of some of the languages, it is likely that some short texts are written equally in some languages. For instance, *Acabo de publicar una foto* (I just published a photo) can be either Spanish or Catalan, and cannot be disambiguated in the absence of more context. This case had to be annotated as *es/ca*.

**Multilingual tweets** contain parts of a tweet in different languages, where the annotators were instructed to annotate all of the languages being used. For instance, *Qeeeee matadaaa* (Spanish: that was exhausting) *da Biyar laneaaaa...* (Basque: and gotta go to work tomorrow) should be annotated as *es+eu*, and *Acho que vi a Ramona hoje* (Portuguese: man, I've seen Ramona today) *but im not sure* (English) should be annotated as *pt+en*. Occasionally, three languages were also found, e.g., *Egun on! Buenos días! Good morning!* (Good morning in Basque, Spanish and English), annotated as *eu+es+en*. The annotation had to consider all the languages being used, in no specific order, except when a single word or term was used as a constituent of a sentence in another language, e.g., *es un outsider* (Spanish: he is an outsider), where only one language is annotated.

The last possible cases are the **mixed tweets**, which are the result of having multilingual tweets where at least one of the languages is either *undeterminable*, *other*, or *ambiguous*. It could also be the case that a multilingual tweet with two languages is the combination two of the cases above, e.g., *other* + *ambiguous*. However, we have not found any of these cases in our dataset. We have only found cases where one of the six languages under study is combined with either *other* or *ambiguous*, which were ultimately removed from the dataset for being very rare and not having enough examples for training, as we describe next.

### 5.3 Annotated Corpus and Evaluation Measures

All the 35,000 tweets were annotated following the aforementioned methodology. Given that the cases where a tweet was annotated as a *mixed tweet* –i.e., where certain language was combined with a language not considered in the task ('lang+other'), or with an ambiguous text ('lang1+lang2/lang3')– were very rare, they were removed from the dataset. These include only 16 cases, which after removing led to an anno-

| Language | Tweets | % |
|---|---|---|
| Spanish (es) | 21,417 | 61.22 |
| Portuguese (pt) | 4,320 | 12.35 |
| Catalan (ca) | 2,959 | 8.46 |
| English (en) | 1,970 | 5.63 |
| Galician (gl) | 963 | 2.75 |
| Basque (eu) | 754 | 2.16 |
| Undeterm. (und) | 787 | 2.25 |
| Multilingual (a+b) | 747 | 2.14 |
| Ambiguous (a/b) | 625 | 1.79 |
| Other | 442 | 1.26 |

**Table 1** Distribution of the manual annotation.

tated corpus composed of 34,984 tweets. The corpus, including also the content of the tweets, can be found on the shared task's web site[6]. Table 1 shows the distribution of the manual annotations, where it can be seen that Spanish is the predominant language, which amounts to 61.22% of the tweets. This is why we use a macroaverage approach to evaluate the systems, as we describe later, which rewards the systems that perform well for all the languages rather than just for the predominant language. Table 2 shows a breakdown of the annotations by region. It shows that the prevalence of Spanish is especially marked in Galicia (86.61%) and in the Basque Country (78.46%). It is more evenly distributed in Catalonia, with 50.62% of the tweets in Spanish and 29.40% in Catalan. Spanish barely occurs in Portugal (only 1.16% of the times), where Portuguese is the predominant language with 81.82% of the tweets. English has a moderate presence across all regions, ranging from 1.55% to 8.28%, and the other three languages –Catalan, Basque, and Galician– have a tiny presence outside their region. The number of ambiguous tweets is much higher in Portugal than in the other regions, especially due to the large number of Portuguese tweets that could also be deemed Galician (pt/gl). Multilingual tweets occur especially in the Basque Country (mostly eu+es), given that code switching occurs very often in this region, and the fact that the two languages are so different makes it easy for the human annotator to identify the presence of both languages; likewise, due to the big difference between both languages, it is very unlikely that a tweet is ambiguous in Spanish or Basque (eu/es). The number of "other" languages is significantly higher in Catalonia than in the other regions, potentially due to the higher diversity of nationalities, due to being a rather touristic region, and a close-by region for the French and Italians, whose languages are considered as "other" in this work.

Additionally, we asked a second annotator for each of the regions to re-annotate a 10% sample of the tweets, i.e., 3,500 tweets altogether. This allows us to compute the inter-annotator agreement on a 10% sample of the whole, so that we can measure the difficulty of the task for human annotators. The inter-annotator agreement is computed as the pairwise agreement between the two annotations for each tweet. Only exact matches are considered as agreement, hence if an annotator labeled a tweet as "gl", and the other annotated it as "es/gl", this is computed as a disagreement. Overall, the annotators agreed 92.6% of the times, distributed by region as shown

[6] http://komunitatea.elhuyar.org/tweetlid/resources/

| | Basque Country | | Catalonia | | Galicia | | Portugal | |
|---|---|---|---|---|---|---|---|---|
| **Language** | **Tweets** | **%** | **Tweets** | **%** | **Tweets** | **%** | **Tweets** | **%** |
| Spanish (es) | 7842 | 78.46 | 5057 | 50.62 | 8460 | 84.61 | 58 | 1.16 |
| Portuguese (pt) | 22 | 0.22 | 44 | 0.44 | 163 | 1.63 | 4091 | 81.82 |
| Catalan (ca) | 20 | 0.20 | 2937 | 29.40 | 1 | 0.01 | 1 | 0.02 |
| English (en) | 595 | 5.95 | 827 | 8.28 | 155 | 1.55 | 393 | 7.86 |
| Galician (gl) | 2 | 0.02 | 2 | 0.02 | 959 | 9.59 | 0 | 0.00 |
| Basque (eu) | 751 | 7.51 | 2 | 0.02 | 1 | 0.01 | 0 | 0.00 |
| Undeterm. (und) | 233 | 2.33 | 386 | 3.86 | 34 | 0.34 | 134 | 2.68 |
| Multilingual (a+b) | 430 | 4.30 | 230 | 2.30 | 40 | 0.40 | 47 | 0.94 |
| Ambiguous (a/b) | 65 | 0.65 | 137 | 1.37 | 167 | 1.67 | 256 | 5.12 |
| Other | 35 | 0.35 | 368 | 3.68 | 19 | 0.19 | 20 | 0.40 |

**Table 2** Distribution of the manual annotation by region.

| Region | Agreement | Most frequent errors |
|---|---|---|
| Basque Country | 93.6% (936/1000) | es → en+es (1.2%, 12 tweets)<br>es+eu → eu (1.0%, 10 tweets)<br>es+eu → es (0.5%, 5 tweets)<br>en+es → es (0.5%, 5 tweets) |
| Galicia | 88.1% (881/1000) | gl → es (4.2%, 42 tweets)<br>en → es (1.9%, 19 tweets)<br>und → es (1.9%, 19 tweets)<br>es/gl → es (1.0%, 10 tweets)<br>es → gl (1.0%, 10 tweets)<br>es → es/gl (0.5%, 5 tweets) |
| Catalonia | 96.0% (960/1000) | es → ca/es (0.5%, 5 tweets) |
| Portugal | 93.0% (465/500) | pt → gl/pt (1.2%, 6 tweets)<br>gl/pt → pt (1.2%, 6 tweets) |

**Table 3** Inter-annotator agreement values distributed by region, computed as the pairwise agreement between two annotators for 10% of the corpus. The last column of the table shows the most frequent disagreements between annotators, where the original annotator picked the value on the left of the arrow, and the second annotator picked the value on the right of the arrow.

in Table 3. These values show that, to some extent, the distinction between similar languages as well as very frequent linguistic interferences can make it difficult for the human annotator. This can be observed especially in the case of Galicia, where the inter-annotator agreement rate is lower than for the other regions. The low inter-annotator agreement values between "es" and "gl" in Galicia can be explained by two factors: on the one hand, the official Galician language uses the same spelling system as Spanish and, on the other hand, the colloquial Galician language often contains many Spanish interferences since people tend to make use of informal Spanish words and expressions. This makes the distinction between the two languages an even more challenging task for human annotators. It is also worth mentioning that while the annotation work for each region will mostly include tweets involving the two languages spoken in the region, there are multiple combinations of those (e.g., es, gl, es+gl, es/gl), besides the fact that other languages also occasionally occur.

Moreover, we also wanted to look at two more factors that are key in our research goals: (1) the length of tweets, to check whether the brevity also makes it more difficult for human annotators, and (2) the fact that tweets are monolingual or

| Tweet length | Agreement | Most frequent errors |
|---|---|---|
| 121-140 | 93.1% (162/174) | en → es (1.72%, 3 tweets) |
| 101-120 | 89.1% (164/184) | es → en+es (5.43%, 10 tweets)<br>gl → es (2.17%, 4 tweets)<br>es+eu → eu (1.63%, 3 tweets) |
| 81-100 | 95.4% (271/284) | en → es (1.06%, 3 tweets)<br>es → gl (1.06%, 3 tweets)<br>gl → es (1.06%, 3 tweets) |
| 61-80 | 96.6% (453/469) | gl → es (1.49%, 7 tweets)<br>es → gl (0.64%, 3 tweets) |
| 41-60 | 96.1% (706/735) | gl → es (0.68%, 5 tweets)<br>es+eu → eu (0.68%, 5 tweets) |
| 21-40 | 91.9% (845/919) | gl → es (2.07%, 19 tweets)<br>en → es (0.87%, 8 tweets)<br>und → es (0.44%, 4 tweets) |
| 1-20 | 81.9% (376/459) | und → es (3.05%, 14 tweets)<br>es/gl → es (1.74%, 8 tweets) |

**Table 4** Inter-annotator agreement values by tweet length. The last column of the table shows the most frequent disagreements between annotators, where the original annotator picked the value on the left of the arrow, and the second annotator picked the value on the right of the arrow.

| Monolingual/multilingual | Agreement | Most frequent errors |
|---|---|---|
| Monolingual | 94.1% (2816/2994) | gl → es (1.40%, 42 tweets)<br>en → es (0.67%, 20 tweets)<br>und → es (0.63%, 19 tweets)<br>es/gl → es (0.33%, 10 tweets)<br>es → gl (0.33%, 10 tweets)<br>es+eu → eu (0.33%, 10 tweets) |
| Multilingual | 60.9% (42/69) | es → en+es (20.29%, 14 tweets) |

**Table 5** Inter-annotator agreement values for monolingual and multilingual tweets. The last column of the table shows the most frequent disagreements between annotators, where the original annotator picked the value on the left of the arrow, and the second annotator picked the value on the right of the arrow.

multilingual. Table 4 shows the agreement values broken down by length. The agreement rates show that there is no significant difference for tweet lengths ranging from 21 to 140 characters. However, the agreement rate drops for tweets between 1 and 20 characters; a number of these cases where due to the difficulty of distinguishing whether a short tweet is written in a certain language or is instead undeterminable, while other cases include confusions between Galician and Spanish or Portuguese, as well as English with Spanish, e.g., due to barbarisms. On the other hand, Table 5 shows the agreement values for monolingual and multilingual tweets. In this case, the agreement rate is substantially lower for multilingual tweets than it is for monolingual tweets. The errors when annotating multilingual tweets include a majority of cases where an annotator labeled a tweet as being only Spanish, while the other labeled it as being in both Spanish and English; again, this depends on each annotator's judgment on whether an English word in a Spanish sentence is a barbarism, or can be considered as a constituent word in Spanish.

The manual annotation work was performed separately for each region, especially given that this facilitates the annotators' work, and it does not require proficient knowledge of the six languages under study. The shared task, however, puts together

all the tweets from the four regions, where the language identifiers need to identify all the languages in the same task.

For the purposes of the shared task, the corpus was split into two random sets of tweets: a training set with 14,991 tweets, and a test set with 19,993 tweets. However, due to restrictions on the use of the Twitter API[7], we distributed the corpora to the participants by including only the tweet IDs. We also provided them with a script to download the content of the tweets having the IDs, which scrapes the web page of each tweet to retrieve the content.

Once the participation period ended we checked the set of tweets in the test set that were still available at the moment. This was done specifically on the 7th of July, with the submission deadlines closed for all the participants. This final check found that 18,423 out of the initial 19,993 tweets, i.e., 92.1%, were available at the moment. For further details into the composition of the corpora, Table 6 shows the distribution of categories for the train and test datasets.

While the reduction of the evaluation dataset to the 92.1% subset was inevitable at the time the shared task took place, the most recent Terms of Service introduced by Twitter allow us to release the content of the tweets along with the dataset. The fact that new tweets may continue to disappear from Twitter's API does no longer affect to the entirety of the dataset then, and will enable additional research using the original dataset. In order to be able to compare results with those of the participants of the task, we also release the list of 18,423 tweet IDs we used for evaluation.

The participants had to submit their results formatted as 'tweet' and 'lang' pairs, referring to the language each tweet in the test set is written. To be considered a valid response, 'lang' can take one of the following forms:

- 'lang1': single language. Possible values are: [es, en, gl, ca, eu, pt, und, other]
- 'lang1+lang2[+lang3]': multiple languages. Any combination of the aforementioned codes are allowed.

It is important to note that 'lang1/lang2[/lang3]' was not a valid answer. If such notation was found, only the first language was taken into account.

When using multiple languages, ('lang1+lang2[+lang3]') a maximum number of 3 languages could be included. If in any case more were provided, the first 3 languages will be taken into account.

### *5.3.1 Evaluation Measures*

The fact that the corpora (as well as the reality of Twitter itself) is imbalanced, where some languages are far more popular than others, is an important issue to be considered when defining the evaluation measures. Besides, given that the language identification task has been defined as a classification problem where tweets can be either multilingual, with more than a language per tweet, or ambiguous, where it is not possible to disambiguate among a set of target languages, the evaluation measures need to be carefully defined to take these into account.

[7] https://dev.twitter.com/terms/api-terms

| Language | %Tweets Train | %Tweets Test |
|---|---|---|
| Spanish (es) | 57.11 (8,562) | 64.02 (11,794) |
| Portuguese (pt) | 14.35 (2,151) | 10.55 (1,943) |
| Catalan (ca) | 9.78 (1,466) | 7.79 (1,435) |
| English (en) | 6.66 (999) | 4.97 (914) |
| Galician (gl) | 3.38 (507) | 2.30 (423) |
| Basque (eu) | 2.53 (380) | 1.94 (358) |
| Undeterm. (und) | 1.25 (188) | 3.01 (555) |
| Multilingual (a+b) | 2.47 (371) | 1.93 (356) |
| Ambiguous (a/b) | 2.31 (346) | 1.41 (260) |
| Other | 0.14 (21) | 2.09 (385) |

**Table 6** Distribution of the manual annotation in train and test data sets.

To deal with the imbalance, we compute the precision, recall, and F1 values for each language, and the macroaveraged measures for all languages afterwards. This is intended to provide higher scores to systems that perform well for many languages, rather than those performing very well in the most popular languages such as Spanish or Portuguese. We compute Precision (P), Recall (F) and F1 measures as defined in Equations 1, 2, and 3.

$$P = \frac{1}{|C|} \sum_{i \in C} \frac{TP_i}{TP_i + FP_i} \tag{1}$$

$$R = \frac{1}{|C|} \sum_{i \in C} \frac{TP_i}{TP_i + FN_i} \tag{2}$$

$$F_1 = \frac{1}{|C|} \sum_{i \in C} \frac{2 \cdot TP_i}{2 \cdot TP_i + FP_i + FN_i} \tag{3}$$

where $C = \{ca, en, es, eu, gl, pt, amb, und\}$ is the set of labels defined in our classification task, and $TP$, $FP$ and $FN$ refer to the counts of true positive, false positive and false negative answers respectively.

The evaluation of our task needs to deal with a ground truth which is occasionally multi-label, so that traditional approaches used in language identification tasks for computing $TP$, $FP$ and $FN$ are not directly applicable. For this purpose, we adapt a concept-based evaluation methodology for multi-label classification [53] to the specific purposes of the task, which we further describe next. To determine whether a system's output for a tweet is correct, we compare it with the manually annotated ground truth. Given that tweets are not simply multilingual, the $TP$, $FP$ and $FN$ values are computed as follows:

– For monolingual tweets, the $TP$ count is incremented by 1 if the answer is correct, and $FP$ is incremented by 1 for the language output by the system otherwise. If a system's prediction contains more than one language, incorrect languages will be penalized, e.g., for a tweet annotated as "pt", a system that outputs "pt+en" will increment $TP$ for "pt" but also $FP$ for "en". $FN$ will be incremented for the language in the ground truth if the answer does not contain the correct language. Hence, the system that outputs "eu" for a tweet that is actually "pt", will count as an additional $FP$ for "eu", and as a $FN$ for "pt".

- For multilingual tweets, we apply the same evaluation methodology as for the multilingual tweets above repeatedly for each of the languages in the ground truth, e.g., for a tweet annotated manually as "ca+es", a system that outputs just "ca" will count as $TP$ for "ca" and as $FN$ for "es".
- For ambiguous tweets that could have been written in any of a set of languages, any of the responses in the ground truth is deemed correct, e.g., for a tweet annotated as "ca/es", either "ca" or "es" is deemed correct as a response, counting as $TP$ of the "amb" category in either case. If, instead, the system outputs "pt", which is not among the languages listed in the ground truth of the ambiguous tweet, the evaluation counts as a $FP$ for "pt", and as a $FN$ for "amb".

Finally, note that we merged tweets annotated as "other" or "und" for evaluation purposes. We did not differentiate between them as those are the tweets that need to be ruled out for being out of the scope of the task. If a system determines that a tweet is "other", and the ground truth is "und", or vice versa, it is deemed correct. To facilitate replication of the experiments as well as comparison of performance results, the evaluation script we used to compute the performance scores is also available on the workshop site.

## 6 Shared Task to Test and Validate the Benchmark

The TweetLID shared task consisted of two separate tracks, one being constrained where external resources were not allowed, and the other being unconstrained were the participants could make use of external resources. Out of the initially registered 16 participants, 7 groups submitted their results for either one or both of the tracks. Participants had a 72 hour window to work with the test set and submit up to two results per track. Next, we first summarize the types of approaches that the participants relied on, and further detail the technique used by each of the participants afterwards.

### 6.1 Overview of the Techniques and Resources Employed

The participants relied on very diverse and different techniques in their systems. They employed different classification algorithms, different methods to learn the models for each language, as well as different criteria to determine the languages of a tweet. This diversity of approaches enables us to broaden the conclusions drawn from the analysis of the performance of different systems. One aspect that the participants agreed upon is the need to preprocess tweets by removing some tokens that do not help for the language identification task such as URLs and user mentions, as well as by lowercasing and reducing the repetition of characters, among others.

The participants used different classification algorithms to develop their systems. The classification algorithms used by most participants include Support Vector Machines (SVM), and Naive Bayes, which have proven effective in previous research in language identification for longer texts.

Not all the participants developed multilabel techniques that can deal with multilingual tweets. Only two of them actually did, mostly by defining a threshold that

| TEAM | Classifier | Representation | Ext. Resources | Multiling. |
|---|---|---|---|---|
| **Citius-imaxin** | 1) ranked n-grams<br>2) naive bayes | words & n-grams<br>& suffixes | news corpora | no |
| **RAE** | support vector machines | n-grams | - | yes |
| **UB/UPC<br>/URV** | 1) linear interpolation<br>2) out-of-place measure | n-grams | - | no |
| **IIT-BHU** | n-gram distances | n-grams | - | no |
| **CERPAMID** | n-gram distances | 3-grams | Europarl corpus<br>Wikipedia | no |
| **ELiRF<br>@ UPV** | 1) support vector machines<br>2) Freeling | words & 4-grams | Wikipedia | yes |
| **LYS @ UDC** | TextCat & langid.py<br>& langdetect | - | Yali | no |

**Table 7** Summary of the main characteristics of the systems developed by the participants

determines the languages to be picked for the output when the classifier provides a higher confidence score for them.

Table 7 summarizes the characteristics of the approaches developed by each of the participants.

## 6.2 Brief Description of the Systems

**Citius-imaxin** [17] submitted two different systems to each of the tracks. On the one hand, a system they called Quelingua builds dictionaries of words ranked by frequency for each language. New tweets are categorized by weighing the ranked words in it, as well as specific suffixes that characterize each language. On the other hand, they build another system based on Naïve Bayes, which has proven accurate in recent research. For the unconstrained track, they fed the systems with news corpora extracted from online journals for all six languages. Their systems do not pick more than one language per tweet, hence not dealing with multilingual tweets. Their bayesian system achieved the best performance for the unconstrained track. Moreover, it was the only system in the task that outperformed its constrained counterpart.

**RAE** [58] submitted two systems only to the constrained track. Their systems rely on n-gram kernels of variable length for each language. The best parameters for each kernel were estimated from the results on the unambiguous examples in the training dataset by cross-validation. They then used Support Vector Machines (SVM) to categorize each new tweet. They relied on a decision tree to interpret the output of the one-vs-all SVM approach, and thus deciding whether the confidence values for more than one language exceeded a threshold (multilingual tweet), only one did (monolingual tweet), or none did (undeterminable).

**UB/UPC/URV** [47] submitted one system to each of the tracks. They developed a different type of system in this case for each track. The first system, submitted to the constrained track, makes use of a linear interpolation smoothing method [28] to compute the probabilities of each n-gram to belong to a language, and weigh new tweets using those probabilities. The second system, submitted to the unconstrained track, is an out-of-place approach that builds a ranked list of n-grams for each lan-

guage in the training phase, and compares each new tweet with these ranked lists to find the language that resembles in terms of n-gram ranks.

**IIT-BHU** [66] only submitted a run to the constrained track. They adapted a system that they previously created for other kinds of texts [65], which is a simple language identification system that makes use of n-grams, and based on that created by Cavnar and Trenkle [12], to the context of Twitter. Basically, they integrated a preprocessing module that removes noisy tokens such as user mentions, hashtags, URLs, etc., and then uses a symmetric cross entropy to measure the similarity or distance between each new tweet and the models learned for each language in the training phase.

**CERPAMID** [75] submitted two systems to each of the tracks. They extract n-grams of three characters to represent the tweets, and use three different weighing methods to weigh the n-grams. Then, they give a score to each new tweet for all the languages in the collection using the three weighing schemes, and pick the final language given as output by the system through simple majority voting. As their systems only output one language, they did not develop any solutions to deal with multilingual tweets. For the unconstrained track, they used the Europarl corpus [33] for English, Spanish, and Portuguese, and Wikipedia for Basque, Catalan, and Galician.

**ELiRF @ UPV** [25] submitted two systems to each of the tracks. For the constrained track, the authors made use of a one-vs-all classifier combining method using SVM. The two approaches submitted to the constrained track differ in the way they deal with multilingual tweets: on one of the approaches, they consider each combination of languages as a new category, while in the other approach they defined a threshold so that the output included all the languages for which the SVM classifier returned a higher confidence value. For the unconstrained track, they developed a classifier using SVM, which used Wikipedia to train the system but did not return multilabel outputs, and another classifier using Freeling's language identification component [55], which includes its own models of 4-grams for the languages in the corpus, except for Basque that the authors created themselves. The constrained method that relies on a threshold to pick the languages for the output achieved the best performance for the constrained track.

**LYS @ UDC** [48] submitted two systems to each of the tracks. They used three different classifiers to develop their systems: TextCat [12], langid.py [41], and langdetect [62]. The two different systems they developed for both tracks differ in that one determines the final output by relying on the classifier with higher confidence, while the other determines the output by majority voting. For the unconstrained track, they used the corpus provided with Yali [44]. Their systems return a single language as output, not dealing with multilingual tweets.

### 6.3 Results

Table 8 shows the results for the *constrained* track, and Table 9 shows the results for the *unconstrained* track. The **ELiRF @ UPV** group, with an SVM-based approach that uses 4-grams and words as features, performed best for the constrained track with an F1 of 0.752. In the unconstrained track, **Citius-imaxin** presented the most

| # | TEAM | P | R | F1 |
|---|---|---|---|---|
| 1 | ELiRF @ UPV II | 0.825 | 0.744 | 0.752 |
| 2 | ELiRF @ UPV I | 0.824 | 0.730 | 0.745 |
| 3 | UB/UPC/URV | 0.777 | 0.719 | 0.736 |
| 4 | RAE II | 0.806 | 0.689 | 0.734 |
| 5 | RAE I | 0.811 | 0.687 | 0.733 |
| 6 | Citius-imaxin II | 0.824 | 0.685 | 0.726 |
| 7 | Citius-imaxin I | 0.689 | 0.743 | 0.699 |
| 8 | CERPAMID I | 0.716 | 0.681 | 0.666 |
| 9 | LYS @ UDC I | 0.732 | 0.734 | 0.638 |
| 10 | IIT-BHU | 0.605 | 0.670 | 0.615 |
| 11 | CERPAMID II | 0.704 | 0.578 | 0.605 |
| 12 | LYS @ UDC II | 0.610 | 0.582 | 0.498 |

**Table 8** Performance results for all the submissions to the constrained track, sorted by F1 measure.

accurate system with a very similar F1 value, 0.753, which uses a bayesian classifier with words, n-grams and suffixes as features.

One of the aspects that stands out from the results of the participants is the fact that most of the systems performed better in the constrained track, and the lower performance of their unconstrained counterparts suggests that either the external resources used are not suitable for the task, or they were not properly exploited. Surprisingly, the only unconstrained algorithm outperforming its constrained counterpart was that by Citius-imaxin. This posits an important caveat of the presented systems, which needs to be further studied in the future.

| # | TEAM | P | R | F1 |
|---|---|---|---|---|
| 1 | Citius-imaxin II | 0.802 | 0.748 | 0.753 |
| 2 | ELiRF @ UPV II | 0.737 | 0.723 | 0.697 |
| 3 | ELiRF @ UPV I | 0.742 | 0.686 | 0.684 |
| 4 | Citius-imaxin I | 0.696 | 0.659 | 0.655 |
| 5 | LYS @ UDC I | 0.682 | 0.688 | 0.581 |
| 6 | UB/UPC/URV | 0.598 | 0.625 | 0.578 |
| 7 | LYS @ UDC II | 0.588 | 0.590 | 0.571 |
| 8 | CERPAMID I | 0.694 | 0.461 | 0.506 |
| 9 | CERPAMID II | 0.583 | 0.537 | 0.501 |

**Table 9** Performance results for all the submissions to the unconstrained track, sorted by F1 measure.

Next, we delve into the performance of the different systems, by looking at the results broken down into different aspects, which allows us to carry out a more detailed analysis of their performance. First, we perform an alternative microaveraged evaluation of the systems, to complement the analysis. Then, we show the performance of baseline approaches, and compare them with the performance of the participants of the shared task. We then analyze each system's performance in more detail, by looking at the three main issues that motivated our work, i.e., brevity of tweets, multilingualism, and similarity between languages. Finally, we analyze the errors of the systems to better understand the limitations of the language identification systems.

### *6.3.1 Alternative Microaveraged Evaluation*

For the sake of comparison with the performance reported in other research works, we also show here the microaveraged evaluation of the three best systems in each track. Note that the micro-averaged evaluation favors the overall performance of the systems, regardless of their likely poor performance for some of the languages. Tables 10 and 11 show the microaveraged results for both tracks, with an overall boost in the results for all the contestants. Still, the best results obtained in this shared task are far from the 99.4% accuracy score reported for formal text, or the 92.4% accuracy score reported for microblogs by Carter et al. [10]. However, it is worth mentioning that Carter et al's scores rely on a monolingual tweet language identification task for major languages including Dutch, English, French, German, and Spanish. The fact that TweetLID has introduced multilingual tweets, as well as tweets from underrepresented languages led to slightly lower performances scores of 89.8% accuracy in the best case. Still, this only reflects a 2.6% accuracy loss when compared to Carter et al's best results for tweets.

| # | TEAM | P | R | F1 |
|---|---|---|---|---|
| 1 | ELiRF @ UPV II | 0.891 | 0.886 | 0.889 |
| 2 | ELiRF @ UPV I | 0.897 | 0.880 | 0.888 |
| 3 | Citius-imaxin I | 0.891 | 0.871 | 0.881 |
| 4 | RAE II | 0.884 | 0.869 | 0.877 |
| 5 | RAE I | 0.882 | 0.866 | 0.874 |
| 6 | UB/UPC/URV | 0.887 | 0.852 | 0.869 |
| 7 | CERPAMID I | 0.856 | 0.838 | 0.847 |
| 8 | Citius-imaxin II | 0.847 | 0.828 | 0.837 |
| 9 | CERPAMID II | 0.832 | 0.815 | 0.824 |
| 10 | LYS @ UDC I | 0.807 | 0.790 | 0.798 |
| 11 | IIT-BHU | 0.781 | 0.790 | 0.786 |
| 12 | LYS @ UDC II | 0.653 | 0.639 | 0.646 |

**Table 10** Microaveraged performance results for all the submissions to the constrained track, sorted by F1 measure.

| # | TEAM | P | R | F1 |
|---|---|---|---|---|
| 1 | Citius-imaxin I | 0.898 | 0.878 | 0.888 |
| 2 | ELiRF @ UPV II | 0.839 | 0.854 | 0.847 |
| 3 | ELiRF @ UPV I | 0.820 | 0.802 | 0.811 |
| 4 | Citius-imaxin II | 0.806 | 0.788 | 0.797 |
| 5 | LYS @ UDC II | 0.792 | 0.776 | 0.784 |
| 8 | CERPAMID I | 0.767 | 0.751 | 0.759 |
| 7 | LYS @ UDC I | 0.749 | 0.733 | 0.741 |
| 9 | CERPAMID II | 0.733 | 0.718 | 0.726 |
| 6 | UB/UPC/URV | 0.715 | 0.701 | 0.708 |

**Table 11** Microaveraged performance results for all the submissions to the unconstrained track, sorted by F1 measure.

| System | P | R | F1 |
|---|---|---|---|
| Twitter | 0.457 | 0.498 | 0.463 |
| TextCat | 0.586 | 0.480 | 0.447 |

**Table 12** Performance results of baseline approaches using existing tools and resources, which enable comparison with the submitted systems.

### *6.3.2 Comparison with Baseline Approaches*

Table 12 includes two additional results as baselines that we computed using the following two solutions: (i) Twitter's metadata, which the system itself provides with each tweet, but it does not recognize Basque, Catalan, and Galician, and (ii) TextCat, a state-of-the-art n-gram-based language identification system developed for formal texts, which can deal with the six languages considered in the task. Note that TextCat was run after cleaning up the tweets by removing hashtags, URLs, and user mentions, as well as lower-casing the texts. The low performance of both solutions, with F1 values below 0.5, emphasizes the difficulty of the task, as well as the need for proper alternatives for social media texts.

### *6.3.3 Evaluation with Respect to Unresolved Issues*

In line with our motivation to study three key unresolved issues in language identification, we now delve into the analysis of results by looking into the performance of the systems when it comes to these three aspects separately: (i) performance results by tweet length, (ii) performance results for monolingual and multilingual tweets, and (iii) performance between similar languages by looking at the confusion matrix.

**(i) Evaluation by Tweet Length.** Figure 1 shows the performance of the systems by tweet length. These boxplots enable the visualization of quartiles in the ranked list of performance values; the bottom and top edges represent 0% and 100% percentiles, the bottom and top of the box represent the 25% and 75% percentiles, and the middle line represents the median, which allows to compare the distributions of performances for different tweet lengths. These results clearly show the tendency of language identifiers to classify with substantially higher accuracy the tweets with more than 60 characters; the performance of the systems progressively drops especially for tweets with fewer than 60 characters. The performance is dramatically lower for tweets as short as 20 characters or fewer. While this corroborates the findings in previous works on language identification, it shows that language identifiers can also perform accurately for long tweets. Even though there is still room for improvement with long tweets, the main challenge remains in the correct identification of language for short tweets.

**(ii) Evaluation for Monolingual and Multilingual Tweets.** Figure 2 shows the results that the systems achieved for monolingual and multilingual tweets. As expected, the language identifiers performed substantially worse for multilingual tweets than for monolingual tweets. It is worth mentioning again that only two of the seven participants produced multilingual labels in their outputs, which means that for the other five systems, the evaluation is performed assuming that they will always miss at

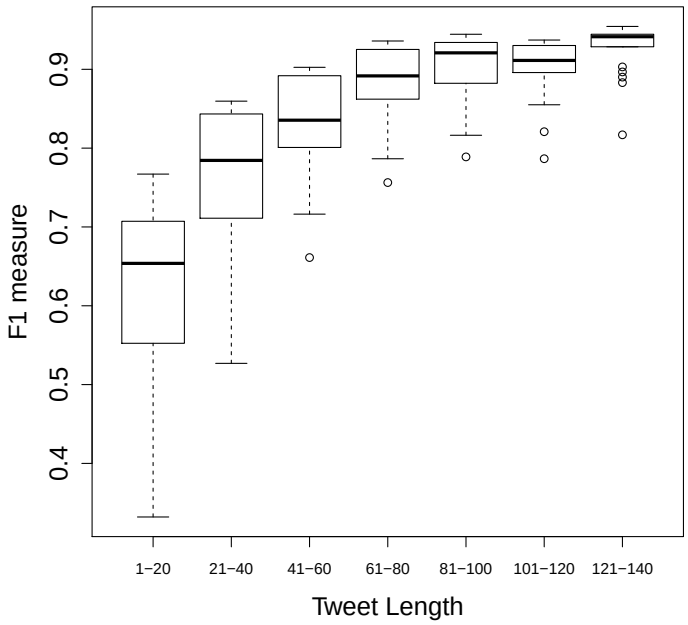


**Fig. 1** F1 scores achieved by submitted systems for different tweet lengths (tweet lengths measured as character counts after removing hashtags, user mentions, and URLs)

least one of the language in the multilingual ground truth. The two systems that produced multilingual labels, ELiRF and RAE, did obtain the best performance scores for the subset of multilingual tweets, with 0.453 and 0.390 F1 measure, respectively. Still, others who did not produce multilingual labels were not far from them, such as IIT-BHU achieving 0.370 F1 measure, and CERPAMID achieving 0.356. Even if the systems who considered multilingualism as a possible output performed better, the relatively small difference with respect to other systems shows the difficulty of dealing with these cases.

Despite the unsurprising fact that the systems performed worse for multilingual tweets, this analysis does, however, help us quantify the difference in terms of F1 measure between monolingual and multilingual tweets, where the classification of the former is about 20% more accurate than the latter. This posits an important drop in performance when tweets are of multilingual nature, which emphasizes the importance of properly dealing with multilingual tweets, and leaves a challenge open for future research in tweet language identification.

**(iii) Evaluation by Language, Focusing on Similar Languages.** Figure 3 summarizes in a boxplot the distribution of precision values achieved by the 21 submitted systems for the different categories. It can be seen that the systems performed poorly especially for Galician (gl); this can be due to its similarity to Spanish (es) and Por-

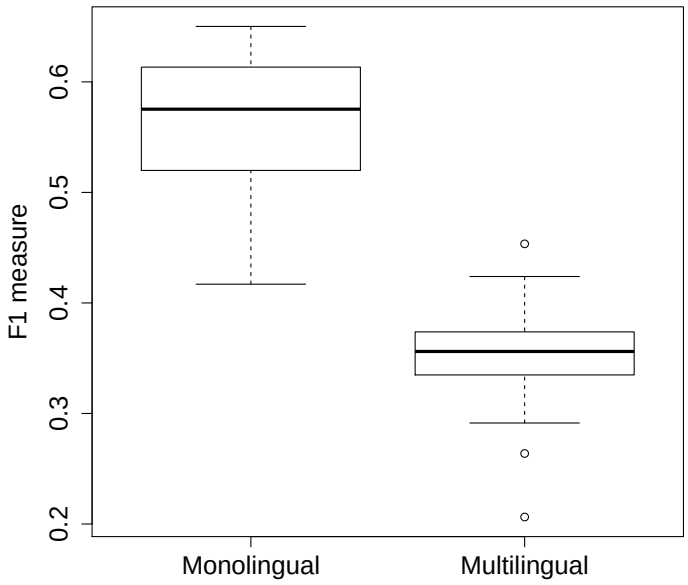


**Fig. 2** F1 scores achieved by the submitted systems for monolingual and multilingual tweets.

tuguese (pt), and its little presence in the corpus. Because of this similarity, and of course the cultural proximity where users tend to mix up spellings, the system might have had a tendency to picking the most popular of the languages in these cases as output. The systems performed better for the rest of the languages, but still surprisingly there is a high variation of performances for Basque (eu), where we can see that some of the systems performed poorly. This is rather surprising given that Basque is very different from the rest of the languages, being an isolate language. A closer look at the errors by the lowest performing systems for Basque shows that these systems have a tendency towards picking the prevalent language (Spanish) for languages that have low representativity in the training set, such as Basque and Galician. Other systems, however, did better in dealing with the imbalance of the data, distinguishing what should be easier to distinguish from the rest of the languages, which is the case of Basque. Galician has, therefore, two challenges, its high similarity with respect to Spanish and Portuguese, as well as the small presence in the training set and the dataset. It also stands out that all the systems performed very well for Spanish, being this the majority language with over 60% of the tweets in the corpora.

Figure 4 complements the analysis with recall values achieved by the systems for the different languages. It can be seen that recall is especially low for undeterminable tweets as well as for Galician tweets. This highlights the difficulty of language identification systems to distinguish these cases from others; in the case of Galician, it

is difficult to distinguish it from Portuguese and Spanish due to their much higher presence, and in the case of undeterminable tweets, it is a challenge to be able to determine that a tweet is not in any of the languages considered by the task, especially because the training set might not have or may have very few tweets in that specific language. Moreover, the recall is also slightly lower for Basque. Even if it is very different from the rest of the languages and hence reasonably easier to identify, its small presence in the training set harms the performance of some of the systems.

Figure 5 enables more detailed visualization of precision and recall values achieved by the systems for Basque and Galician, which as we mentioned above have proven challenging. These two charts show high diversity in the performance of the different systems, with few systems achieving a competitive balance of recall and precision values. The two systems performing best in these two cases, ELiRF for Basque and Citius-imaxin for Galician, have also achieved the best performances for the two tracks of the shared task.

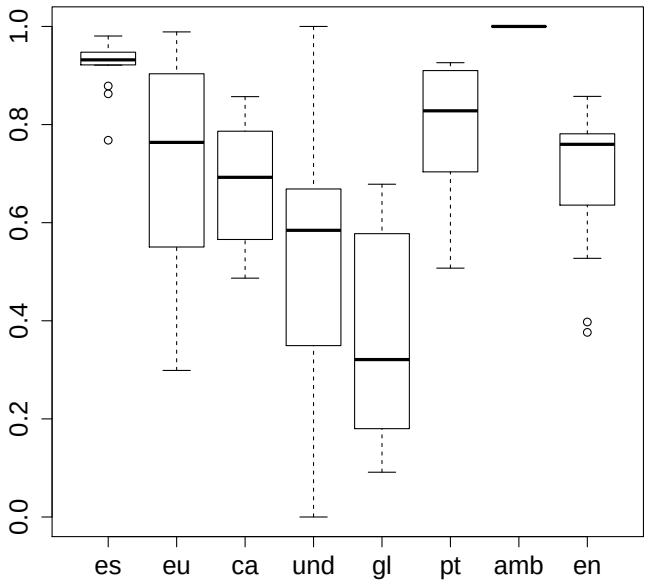


**Fig. 3** Distribution of precision scores by language for the 21 submitted systems, including results for both the constrained and the unconstrained tracks.

Table 13 shows a confusion matrix comparing the ground truth and the aggregated outputs of all the systems for monolingual tweets, which allows us to analyze the extent to which the language identifiers tend to confuse between similar languages. To do this analysis, it is important to consider the bias derived from the skewed distribu-

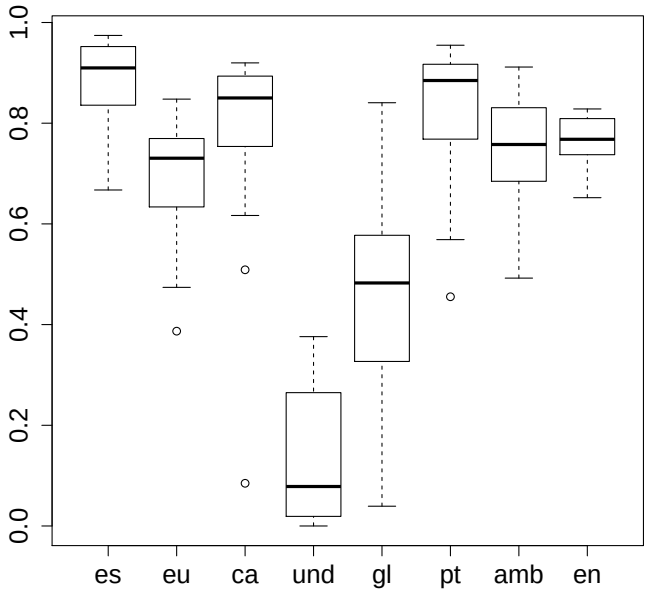


**Fig. 4** Distribution of recall scores by language for the 21 submitted systems, including results for both the constrained and the unconstrained tracks.

tion of tweets (a majority of them in Spanish) in both the training and test datasets. If we do not consider Spanish, Galician language tends to be mostly confused with Portuguese (12.7% errors from the total decisions), which is its closest linguistically related language. Similarly, besides Spanish, Portuguese is confused with Galician (3.2%) more often than with Catalan (1.3%), English (1.1%), or Basque (0.5%). In the case of Spanish, it is mostly confused with the other three Romance languages: Galician (3.5%), Catalan (2.4%), and Portuguese (2.2%), setting aside less related languages, namely English and Basque. Despite this was an anticipated and largely expected outcome, it emphasizes that language similarity is an important issue that reveals the shortcomings of state-of-the-art language identifiers.

### *6.3.4 Misclassified Tweets*

Now we look at the errors produced by the participating systems, as well as the benefit that they could obtain from one another by combining them into a single classifier. First, we combined the output of all the participating systems by majority vote, so we we can obtain a single output for each tweet by aggregating the outputs. Table 14 shows the performance of a system that would combine all the systems, and compares it to that of the best system developed by ELiRF @ UPV. The combined system can

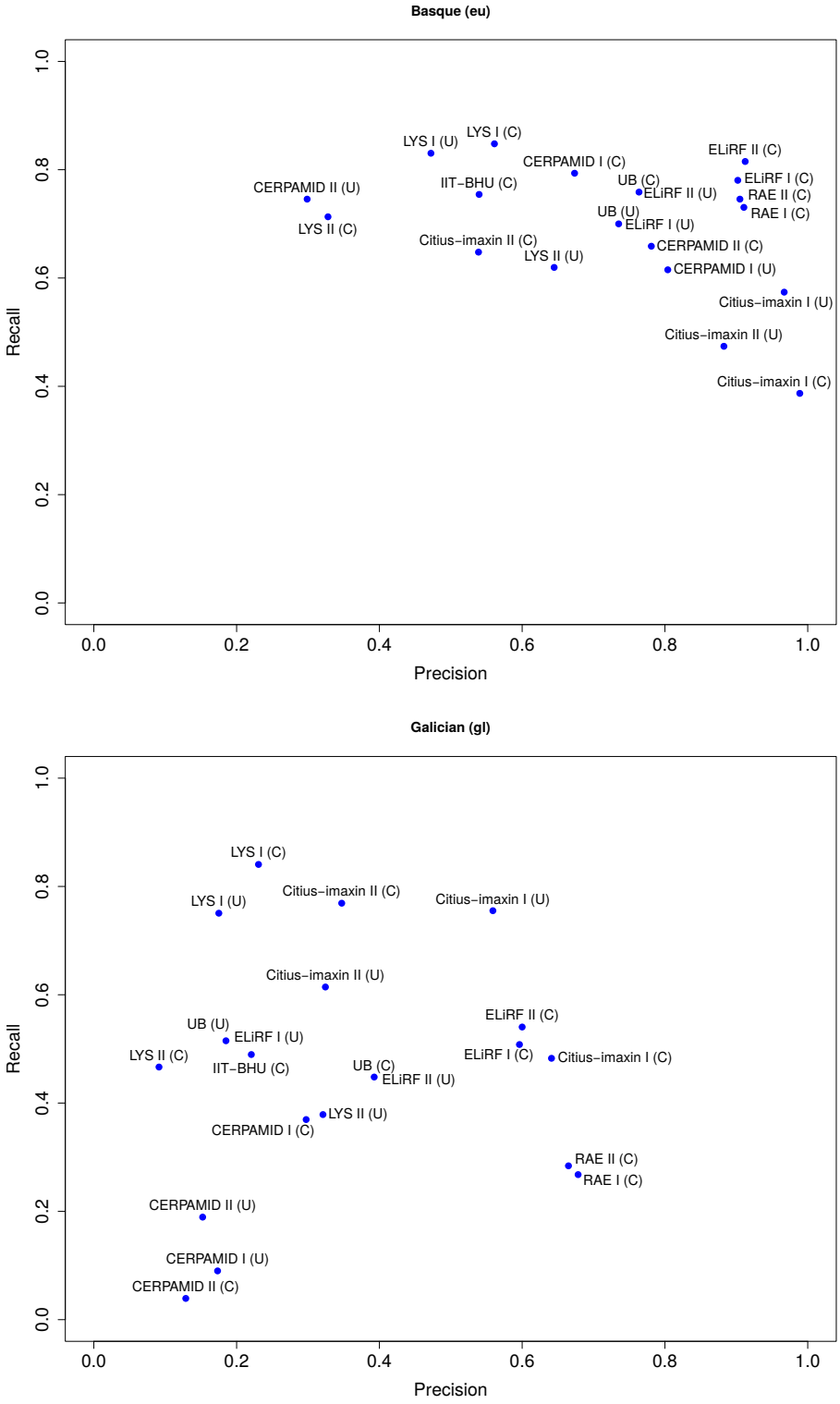


**Fig. 5** Scatter plots showing the precision and recall values for the 21 submitted systems, for tweets in Basque and Galician.

| | pt | ca | es | en | eu | gl | other | multi | und |
|---|---|---|---|---|---|---|---|---|---|
| pt | **83.6** | 1.3 | 8.6 | 1.1 | 0.5 | 3.2 | 0.6 | 0.6 | 0.5 |
| ca | 1.6 | **81.8** | 12.5 | 1.3 | 0.5 | 0.8 | 0.8 | 0.5 | 0.3 |
| es | 2.2 | 2.4 | **88.1** | 1.7 | 0.7 | 3.0 | 0.7 | 0.4 | 0.7 |
| en | 1.1 | 2.9 | 4.6 | **87.4** | 0.9 | 0.8 | 1.0 | 0.5 | 0.8 |
| eu | 0.9 | 1.2 | 10.1 | 2.9 | **76.4** | 1.5 | 1.9 | 1.3 | 3.8 |
| gl | 12.7 | 1.5 | 33.6 | 0.8 | 0.3 | **47.2** | 2.1 | 1.4 | 0.4 |

**Table 13** Confusion matrix showing the percentage of cases whose ground truth is the language in the column and has instead been classified as the language in the row by the systems. The sum of the values in a row add up to 100%. The values in the diagonal (in bold) represent correct classifications, while the rest represent the percentage of deviations from the language in the column to the language in the row.

| Macro-averaged | | | |
|---|---|---|---|
| System | P | R | F |
| Meta-learning | **0.832** | **0.757** | **0.768** |
| Best system (ELiRF @ UPV II) | 0.825 | 0.744 | 0.752 |
| Micro-averaged | | | |
| Meta-learning | **0.910** | **0.892** | **0.901** |
| Best system (ELiRF @ UPV II) | 0.891 | 0.886 | 0.889 |

**Table 14** Results of a meta-learning approach that combines the output of all participating systems by majority vote, compared with the results for the best system in the shared task.

outperform the best system, with slightly better results when both macro-averaging or micro-averaging the performance values.

Among the 18,423 tweets considered for evaluation in the test set, we identified 600 tweets that were not guessed correctly by any of the submitted systems. Next, we look at some of these tweets, which allows us to analyze examples of the most challenging cases.

**Multilingual tweets with low presence of one of the languages.** This kind of tweets are probably the most difficult to deal with because both user intent and cultural habits are combined. Code-switching phenomena are a constant on social media as we have observed in this work. These tweets can often present a challenge even for human annotators. If we take a look at example 1, the use of the verb in Catalan (ets = you're) denotes the intent of the user to write in Catalan. The second part however, is written in Spanish (lo mejorcito = the best). In example 2 the tweet is mainly written in Basque, but the writer ends with a Spanish expression (si o si = come what may).

**Example 1** @username ets lo mejorcito
*[most systems categorized as "es", while it actually combines "ca+es"]*

**Example 2** Dutxita eta gerrate zibilakin gaur bukatubiou si o si
*[most systems categorized as eu, but it actually combines "es+eu"]*

**Lack of identification of NEs.** In some cases, tweets written in a certain language contain NEs which are written in their original language. In the example below, even though the tweet is written in Galician, it contains the name of a TV show in Spanish ("Hay una cosa que te quiero decir"). Not identifying the NE leads the systems to confusion.

**Example 3** Ese neno de “Hay una cosa que te quiero decir” é puuro amorr
*[most systems categorized as “es”, while it is actually “gl”, but the NE (quoted) is indeed “es” and confuses the classifiers]*

**Difficulty to identify undeterminable (“und”) tweets.** In some cases, due to lack of clarity, or because of the brevity of some tweets, not even a human can determine what language is used in a tweet. For instance, in the examples 4 and 5, it is hard to determine what the meaning of the tweets is without additional context, which led to the manual annotation as undeterminable (“und”). It is a challenge, though, for a language identifier to realize what these cases are. The systems generated very diverse predictions for these cases, suggesting that there is no strong similarity with any of the languages.

**Example 4** *@user skiada top!*

**Example 5** *Tu + eu = uiui*

**Difficulty to identify tweets written in other languages.** The correct prediction on these cases should be “other”, as the tweets are written in a language different from those considered in the task. Examples 6 and 7 show two tweets written in Dutch. Participant systems had no language model for that language, and therefore were unable to determine what language it was, and even to determine that it is not one of the languages under consideration. In general, our intuition is that this kind of tweets obtain similarity scores that suggest that they are not far enough from the other language models so as to be regarded as “other”.

**Example 6** #CaminoVascoDelInterior We zijn in Spanje : eerste mojn met schelp in Irún {URL}
*[the systems generated very different predictions, while a tweet in Dutch should be marked as “other”]*

**Example 7** Naar bed naar bed zij duimelot
*[the systems generated very different predictions, while a tweet in Dutch should be marked as “other”]*

The list above summarizes the most frequent types of categorization errors when we look at the tweets misclassified by all of the systems analyzed in this work. Other common errors, such as deviations between similar languages, do not appear in this list given that they are usually guessed correctly by at least one of the systems.

## 7 Discussion

In this work, we describe and release a benchmark dataset and evaluation framework for tweet language identification. Through the shared task we organized to encourage researchers to submit the results of their language identification systems applied to this dataset, we looked at content-based tweet language classification approaches. The study of other features that a social network like Twitter can offer, such as user metadata, are not within the scope of this work and are left for future work.

### 7.1 Performance of Tweet Language Identification Systems

We were especially interested in this case in studying state-of-the-art approaches for language identification in a new scenario like Twitter. However, we do believe that the use of features inherent to the social network can be of help for a language identifier, especially for adding context when the content is insufficient. We believe that the study of additional user-related features can help (i) when tweets are very short, looking for instance at previous tweets posted by a user, which might reveal what language the user uses most, and (ii) when two similar languages need to be distinguished, for instance looking at the location of a user, which might help identify the language(s) that are likely used in that location.

As we have shown, multilingualism is also a challenging issue in short texts like tweets. Further exploiting the social network, one could look at the historical tweets of a user to first list the languages that a user is likely to use, to then determine if the user has used a combination of those in new tweets; this involves having to look at more tweets from a user though, which is costly in terms of API accesses required, and might not always be feasible.

Regarding NEs, none of the participants tried to incorporate NER capabilities to their system, but it could have been useful as shown in some of the misclassified examples above. However, the use of NEs for this task is not trivial. For the shared task, our choice was to ignore NEs when annotating the language. While some NEs can be good hints about the language of the user, such as place names because they are usually translated into the corresponding language (e.g., Donostia (eu) vs. San Sebastian (es), other NEs however tend to be used both in their original form and in their translated form, e.g., Spanish tweeters use both “Game of Thrones” (en) and “Juego de Tronos” (es).

While the shared task we conducted, as well as the analysis of the submitted systems we discuss here, do not consider other social network features beyond a tweet’s content, the dataset we created and released to the scientific community does allow to collect and incorporate these extra features for further analysis.

### 7.2 Comparing Errors between Human Annotators and by Language Identification Systems

Throughout the paper, we have studied both the performance of human annotators as well as that of the automatic language identification systems. The human annotations have been assessed by having two annotators annotate a 10% sample of the whole, while the automatic systems have been evaluated comparing against the manually defined annotation as the ground truth. Both evaluations have shown, to some extent, a similar tendency; both humans and systems struggled to identify the language of short tweets as well as the languages in multilingual tweets, and also found it difficult to distinguish similar languages. Still, there are a number of differences between the performances of humans and systems, which helps us set forth a set of objectives for future work.

**Length of tweets:** while human annotators performed lower for very short tweets of less than 20 characters, the performance was quite consistent for other lengths. The systems, though, showed a progressive decay in performance as the tweets are shorter, experiencing a significant drop in performance for tweets of less than 60 characters. While improving the performance of the language identification for tweets of less than 20 characters might not be viable, we believe that there is still room for improvement for tweets between 20 and 60 characters, which humans could label as accurately as longer tweets.

**Multilingualism:** multilingual tweets have proven challenging both for human annotators and for language identification systems, with a significantly lower performance than for monolingual tweets. However, only two of the participants in the shared task developed systems that would ever output a multilingual label, which makes our analysis in this aspect still inconclusive enough so as to conclude the extend to which it can be improved. The better performance of the two systems that implemented multilingual outputs over the rest of the systems, however, does encourage to perform further research. We believe that testing more multilingual systems would help extend the analysis of classifying multilingual tweets.

**Similar languages:** the confusion between similar languages occurred differently for human annotators, given that each annotator had to deal only with tweets from a specific region, which means that there could be rarely confusions between Spanish and Portuguese, because they usually appear in different regions. Still, for one of the most common errors in our dataset, i.e., confusions between Galician and Spanish, human annotators performed much better, and language identification systems missed as many as 33.6%. In the latter case, the performance worsens owing to the fact that Galician has fewer instances in the training set, which also occurred with Basque, a language which is very different from the rest, but its low presence occasionally harms the performance of classifiers. Better dealing with similar languages, as well as better managing languages with fewer instances in the training set, are certainly two of the key aspects to look at in the future.

### 7.3 Contributions and Limitations of the Shared Task

Through the organization of TweetLID as a shared task, we have fulfilled most of the objectives we set forth at the beginning of planning this work, and we expect that our contributions will help pave the way to researchers aiming to study tweet language identification in the future. However, we have also identified a set of limitations in the shared task.

On the positive side, we believe that TweetLID has managed to attract a good number of participants, who have submitted a diverse set of systems. This has enabled a quite complete analysis of language identification systems applied to tweets, as well as the identification of main directions for future research. This has been possible thanks to the creation of an annotated corpus of tweets that meet the main characteristics we sought, as well as the definition of the evaluation methodology. This corpus will in turn enable further research in the future. Thankfully, Twitter's newly revised terms of service allows us to release the content and all the metadata

of the tweets, which will guarantee that whoever is interested will be able to retrieve the complete dataset, which will not shrink over time.

On the other hand, one of the weak points of the systems submitted to the shared task has been the limited attempt at dealing with multilingual tweets. In fact, only two of the seven participants produced language identification systems that would ever return a multilingual label as output. While it has not been possible to test additional multilingual systems in this shared task, it would have been useful to have more such systems participating, and would be ideal to have in a future shared task. Moreover, even if it was originally restricted in the definition of the shared task, we have not let participants to make use of tweet metadata to identify languages, which would also be wise to study in an upcoming shared task.

Last but not least, it also makes it extra challenging to organize the shared task the fact that Twitter's terms of service did not allow us to share tweet content with the participants. Instead, we gave them the list of tweet IDs, which they used to retrieve the content of the tweets themselves by accessing Twitter's API, which leads to each participants having slightly different sets of tweets due to some tweets becoming unavailable over time. The updated terms of service would enable, however, to share the content with the participants of future tasks.

## 8 Conclusion

The Twitter dataset with nearly 35,000 tweets with language label manually annotated has enabled us to study currently unresolved issues in language identification. These include the following three issues: (i) short texts provide very little context to determine the language of their content, (ii) multilingual texts make it more difficult identify the presence of the different languages, and (iii) similar languages are very difficult to distinguish from each other. Our Twitter dataset provides a suitable resource to study the aforementioned issues, which we have put into practice and analyzed through the TweetLID shared task where seven participants submitted the output of their language identifiers. Thanks to the development of this dataset and the shared task to assess the performance of different systems, we have come up with an evaluation methodology that can be of help to researchers in the field.

Our dataset included the five top languages of the Iberian Peninsula –Spanish, Portuguese, Catalan, Basque, and Galician– as well as English. This has allowed participants to compare their systems with four romance languages that share similarities with one another, and two more languages that are substantially different from the rest, i.e., English and Basque. The participants have applied state-of-the-art language identification techniques designed for other kinds of texts such as news articles, as well as adapted approaches that take into account the nature of the brevity and chatspeak found in tweets. Still, the performance of the systems posits the need of further research to come up with more accurate language identification systems for social media. Some of the key shortcomings that the shared task has brought to light include the need for a better choice of external resources to train the systems, the low accuracy of the systems when dealing with underrepresented languages which are very similar to others –as occurred with Galician here–, and the inability to identify

multilingual tweets. Future work on tweet language identification should look into these issues to develop more accurate systems.

**Acknowledgements** This work has been supported by the following projects: PHEME FP7 project (grant No. 611233), QTLeap FP7 project (grant No. 610516), Spanish MICINN projects *Tacardi* (Grant No. TIN2012-38523-C02-01) and *Skater* (Grant No. TIN2012-38584-C06-01), Galician HPCPLN project (Grant No. EM13/041), Celtic (Innterconecta program, Grant No. 2012-CE138).

## References

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proceedings of the Workshop on Languages in Social Media, pp. 30–38. Association for Computational Linguistics (2011)
2. Alegria, I., Aranberri, N., Comas, P.R., Fresno, V., Gamallo, P., Padró, L., San Vicente, I., Turmo, J., Zubiaga, A.: Tweetnorm_es corpus: an annotated corpus for spanish microtext normalization. In: Proceedings of the Language Resources and Evaluation Conference (2014)
3. Baldwin, T., Lui, M.: Language identification: The long and the short of the matter. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 229–237. Association for Computational Linguistics (2010)
4. Baykan, E., Henzinger, M., Weber, I.: Web page language identification based on urls. Proceedings of the VLDB Endowment **1**(1), 176–187 (2008)
5. Beesley, K.R.: Language identifier: A computer program for automatic natural-language identification of on-line text. In: Proceedings of the 29th Annual Conference of the American Translators Association, vol. 47, p. 54. Citeseer (1988)
6. Bergsma, S., McNamee, P., Bagdouri, M., Fink, C., Wilson, T.: Language identification for creating language-specific twitter collections. In: Workshop on Language in Social Media, pp. 65–74. ACL (2012)
7. Brown, R.D.: Finding and identifying text in 900+ languages. Digital Investigation **9**, S34–S43 (2012)
8. Brown, R.D.: Selecting and Weighting NGrams to Identify 1100 Languages. In: Text, Speech, and Dialogue, pp. 475–483 (2013)
9. Cárdenas-Claros, M., Isharyanti, N.: Code-switching and code-mixing in internet chatting: Between'yes,"ya,'and'si'-a case study. The Jalt Call Journal **5**(3), 67–78 (2009)
10. Carter, S., Weerkamp, W., Tsagkias, M.: Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. Language Resources and Evaluation **47**(1), 195–215 (2013)
11. Cassidy, T., Ji, H., Ratinov, L.A., Zubiaga, A., Huang, H.: Analysis and enhancement of wikification for microblogs with context expansion. In: Proceedings of COLING, the 24th International Conference on Computational Linguistics, vol. 12, pp. 441–456 (2012)
12. Cavnar, W.B., Trenkle, J.M., et al.: N-gram-based text categorization. Ann Arbor MI **48113**(2), 161–175 (1994)
13. Chepovskiy, A., Gusev, S., Kurbatova, M.: Language identification for texts written in transliteration. CDUD 2012–Concept Discovery in Unstructured Data p. 13 (2012)
14. Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., Bontcheva, K.: Analysis of named entity recognition and linking for tweets. Information Processing & Management **51**(2), 32–49 (2015)
15. Druck, G.: Generalized expectation criteria for lightly supervised learning. Ph.D. thesis, University of Massachusetts Amherst (2011)
16. Dunning, T.: Statistical identification of language. Computing Research Laboratory, New Mexico State University (1994)
17. Gamallo, P., Garcia, M., Sotelo, S., Pichel, J.R.: Comparing ranking-based and naive bayes approaches to language detection on tweets. In: TweetLID@SEPLN (2014)
18. Gella, S., Bali, K., Choudhury, M.: "ye word kis lang ka hai bhai?" testing the limits of word level language identification. NLPAI (2014)
19. Goldszmidt, M., Najork, M., Paparizos, S.: Boot-strapping language identifiers for short colloquial postings. In: Machine Learning and Knowledge Discovery in Databases, pp. 95–111. Springer (2013)

20. Gottron, T., Lipka, N.: A comparison of language identification approaches on short, query-style texts. In: Advances in information retrieval, pp. 611–614. Springer (2010)
21. Grefenstette, G.: Comparing two language identification schemes (1995)
22. Guo, S., Chang, M.W., Kiciman, E.: To link or not to link? a study on end-to-end tweet entity linking. In: HLT-NAACL, pp. 1020–1030 (2013)
23. Hammarström, H.: A FineGrained Model for Language Identification. In: Proceedings of Improving Non English Web Searching (iNEWS'07), pp. 14–20 (2007)
24. Hughes, B., Baldwin, T., Bird, S., Nicholson, J., MacKinlay, A.: Reconsidering language identification for written language resources (2006)
25. Hurtado, L.F., Pla, F., Giménez, M., Sanchis, E.: Elirf-upv en tweetlid: Identificación del idioma en twitter. In: TweetLID@SEPLN (2014)
26. Ingle, N.: A language identification table. Technical translation international (1980)
27. Jehl, L., Hieber, F., Riezler, S.: Twitter translation using translation-based cross-lingual retrieval. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, pp. 410–421. Association for Computational Linguistics (2012)
28. Jelinek, F.: Statistical methods for speech recognition. MIT press (1997)
29. Kaufmann, M., Kalita, J.: Syntactic normalization of twitter messages. In: International conference on natural language processing, Kharagpur, India (2010)
30. Keesan, C.: Identification of written slavic languages. In: Proceedings of the 28th Annual Conference of the American Translators Association, pp. 517–528 (1987)
31. Kikui, G.i.: Identifying, the coding system and language, of on-line documents on the internet. In: Proceedings of the 16th conference on Computational linguistics-Volume 2, pp. 652–657. Association for Computational Linguistics (1996)
32. King, B., Abney, S.P.: Labeling the languages of words in mixed-language documents using weakly supervised methods. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies, pp. 1110–1119 (2013)
33. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT summit, vol. 5, pp. 79–86 (2005)
34. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: The good the bad and the omg! In: Proceedings of the International Conference on Weblogs and Socila Media, pp. 538–541 (2011)
35. Laboreiro, G., Bošnjak, M., Sarmento, L., Rodrigues, E.M., Oliveira, E.: Determining language variant in microblog messages. In: Proceedings of the 28th ACM/SIGAPP Symposium On Applied Computing, pp. 902–907. ACM (2013)
36. Lehman, B.: The evolution of languages on twitter. `http://blog.gnip.com/twitter-language-visualization/` (2014)
37. Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., Lee, B.S.: Twiner: named entity recognition in targeted twitter stream. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 721–730. ACM (2012)
38. Ljubešić, N., Mikelić, N., Boras, D.: Language indentification: How to distinguish similar languages? In: Proceedings of the 29th International Conference on Information Technology Interfaces, pp. 541–546. IEEE (2007)
39. Lui, M., Baldwin, T.: Multilingual language identification: Altw 2010 shared task dataset. In: Australasian Language Technology Association Workshop 2010, p. 4 (2010)
40. Lui, M., Baldwin, T.: Cross-domain feature selection for language identification. In: In Proceedings of 5th International Joint Conference on Natural Language Processing. Citeseer (2011)
41. Lui, M., Baldwin, T.: langid. py: An off-the-shelf language identification tool. In: Proceedings of ACL, pp. 25–30. ACL (2012)
42. Lui, M., Baldwin, T.: Accurate language identification of twitter messages. In: Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM), pp. 17–25. Association for Computational Linguistics, Gothenburg, Sweden (2014)
43. Lui, M., Lau, J.H., Baldwin, T.: Automatic detection and language identification of multilingual documents. Transactions of the Association for Computational Linguistics **2**, 27–40 (2014)
44. Majliš, M.: Yet another language identifier. In: Student Research Workshop at EACL'12, pp. 46–54. ACL (2012)
45. Martins, B., Silva, M.J.: Language identification in web pages. In: Proceedings of SAC, pp. 764–768. ACM (2005)
46. McNamee, P.: Language identification: A solved problem suitable for undergraduate instruction. Journal of Computing Sciences in Colleges **20**(3), 94–101 (2005)

47. Mendizabal, I., Carandell, J., Horowitz, D.: Tweetsafa: Tweet language identification. In: TweetLID@SEPLN (2014)
48. Mosquera, Y.D., Vilares, D., Vilares, J.: Identificación automática del idioma en twitter: Adaptación de identificadores del estado del arte al contexto ibérico. In: TweetLID@SEPLN (2014)
49. Murthy, K.N., Kumar, G.B.: Language identification from small text samples. Journal of Quantitative Linguistics **13**(1), 57–80 (2006)
50. Myers-Scotton, C.: Contact linguistics: Bilingual encounters and grammatical outcomes. Oxford University Press Oxford (2002)
51. Newman, P.: Foreign language identification: First step in the translation process. Tech. rep., Sandia National Labs., Albuquerque, NM (USA) (1987)
52. Nguyen, D., Doğruöz, A.S.: Word level language identification in online multilingual communication. In: Proceedings of the Conference on Empirical Methods on Natural Language Processing (2014)
53. Nowak, S., Lukashevich, H., Dunker, P., Rüger, S.: Performance measures for multilabel evaluation: a case study in the area of image classification. In: Proceedings of the international conference on Multimedia information retrieval, pp. 35–44. ACM (2010)
54. O'Connor, B., Krieger, M., Ahn, D.: Tweetmotif: Exploratory search and topic summarization for twitter. In: ICWSM (2010)
55. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (2012)
56. Padró, M., Padró, L.: Comparing methods for language identification. Procesamiento del lenguaje natural **33**, 155–162 (2004)
57. Paolillo, J.C.: Conversational codeswitching on usenet and internet relay chat. Herring, Susan C.(ed.) (2011)
58. Porta, J.: Twitter language identification using rational kernels and its potential application to sociolinguistics. In: TweetLID@SEPLN (2014)
59. Prager, J.M.: Linguini: Language identification for multilingual documents. In: Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on, pp. 11–pp. IEEE (1999)
60. Scannell, K.: The Crúbadán Project: Corpus building for underresourced languages. In: Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop, incorporating Cleaneval, vol. 5, p. 5 (2007)
61. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) **34**(1), 1–47 (2002)
62. Shuyo, N.: Language detection library for java (2010)
63. Sibun, P., Reynar, J.C.: Language identification: Examining the issues (1996)
64. Sibun, P., Spitz, A.L.: Language determination: Natural language processing from scanned document images. In: Proceedings of the fourth conference on Applied natural language processing, pp. 15–21. Association for Computational Linguistics (1994)
65. Singh, A.K.: Study of some distance measures for language and encoding identification. In: Workshop on Linguistic Distances, pp. 63–72. ACL (2006)
66. Singh, A.K., Goyal, P.: A language identification method applied to twitter data. In: TweetLID@SEPLN (2014)
67. Tetreault, J., Blanchard, D., Cahill, A.: A report on the first native language identification shared task. In: Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 48–57. Citeseer (2013)
68. Tromp, E., Pechenizkiy, M.: Graph-based n-gram language identification on short texts. In: Proc. 20th Machine Learning conference of Belgium and The Netherlands, pp. 27–34 (2011)
69. Řehůřek, R., Kolkus, M.: Language identification on the web: Extending the dictionary method. In: Computational Linguistics and Intelligent Text Processing, pp. 357–368. Springer (2009)
70. Vatanen, T., Väyrynen, J.J., Virpioja, S.: Language identification of short text segments with n-gram models. In: LREC. Citeseer (2010)
71. Vogel, J., Tresner-Kirsch, D.: Robust Language Identification in Short , Noisy Texts : Improvements to LIGA. In: Proceedings of the 3rd International Workshop on Mining Ubiquitous and Social Environments (MUSE), pp. 1–9. Bristol, UK (2012)
72. Winkelmolen, F., Mascardi, V.: Statistical Language Identification of Short Texts. In: Proceedings of the 3rd International Conference on Agents and Artificial Intelligence, pp. 498–503. Rome, Italy (2011)
73. Xafopoulos, A., Kotropoulos, C., Almpanidis, G., Pitas, I.: Language identification in web documents using discrete hmms. Pattern recognition **37**(3), 583–594 (2004)

74. Xia, F., Lewis, W.D., Poon, H.: Language id in the context of harvesting language data off the web. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 870–878. Association for Computational Linguistics (2009)
75. Zamora, J.D., Bruzón, A.F., Bueno, R.O.: Tweets language identification using feature weighting. In: TweetLID@SEPLN (2014)
76. Zampieri, M.: Using bag-of-words to distinguish similar languages: How efficient are they? In: Computational Intelligence and Informatics (CINTI), 2013 IEEE 14th International Symposium on, pp. 37–41. IEEE (2013)
77. Zubiaga, A., San Vicente, I., Gamallo, P., Pichel, J.R., Alegria, I., Aranberri, N., Ezeiza, A., Fresno, V.: Overview of tweetlid: Tweet language identification at sepln 2014. TweetLID@SEPLN (2014)
78. Zubiaga, A., Spina, D., Amigó, E., Gonzalo, J.: Towards real-time summarization of scheduled events from twitter streams. In: Proceedings of the 23rd ACM conference on Hypertext and social media, pp. 319–320. ACM (2012)