Lang Resources & Evaluation (2013) 47:195–215
DOI 10.1007/s10579-012-9195-y

Microblog language identification: overcoming the limitations of short, unedited and idiomatic text

Simon Carter · Wouter Weerkamp · Manos Tsagkias

Published online: 28 June 2012

Abstract Multilingual posts can potentially affect the outcomes of content analysis on microblog platforms. To this end, language identification can provide a monolingual set of content for analysis. We find the unedited and idiomatic language of microblogs to be challenging for state-of-the-art language identification methods. To account for this, we identify five microblog characteristics that can help in language identification: the language profile of the blogger (blogger), the content of an attached hyperlink (link), the language profile of other users mentioned (mention) in the post, the language profile of a tag (tag), and the language of the original post (conversation), if the post we examine is a reply. Further, we present methods that combine these priors in a post-dependent and post-independent way. We present test results on 1,000 posts from five languages (Dutch, English, French, German, and Spanish), which show that our priors improve accuracy by 5 % over a domain specific baseline, and show that post-dependent combination of the priors achieves the best performance. When suitable training data does not exist, our methods still outperform a domain unspecific baseline. We conclude with an examination of the language distribution of a million tweets, along with temporal analysis, the usage of twitter features across languages, and a correlation study between classifications made and geo-location and language metadata fields.

This work represents a substantially expanded version of Carter et al. (2011) and Weerkamp et al. (2011).

S. Carter (✉) · W. Weerkamp · M. Tsagkias
ISLA, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
e-mail: s.c.carter@uva.nl

W. Weerkamp
e-mail: w.weerkamp@uva.nl

M. Tsagkias
e-mail: e.tsagkias@uva.nl

Keywords Text classification · Language identification · Microblogs · Content analysis

1 Introduction

Microblogging platforms such as Twitter have become important real-time information resources (Golovchinsky and Efron 2010), with a broad range of uses and applications, including event detection (Sakaki et al. 2010; Vieweg et al. 2010), media analysis (Altheide 1996), mining consumer and political opinions (Jansen et al. 2009; Tumasjan et al. 2010), and predicting movie ratings (Oghina et al. 2012). Microbloggers participate from all around the world contributing content, usually, in their own native language. Language plurality can potentially affect the outcomes of content analysis and retrieval of microblog posts (Massoudi et al. 2011), and we therefore aim for a monolingual content set for analysis. To facilitate this, language identification becomes an important and integrated part of content analysis. We address the task of language identification in microblog posts.

Language identification has been studied in the past (see Sect. 2 for previous work in this field), showing successful results on structured and edited documents. Here, we focus on an other type of documents: user generated content, in the form of microblog posts. Microblog posts ("tweets," "status updates," etc.) are a special type of user generated content, mainly due to their limited size, which has interesting effects. People, for example, use word abbreviations or change word spelling so their message can fit in the allotted space, giving rise to a rather idiomatic language that is difficult to match with statistics from external corpora.

Document length has been shown to significantly affect language identification, shorter documents being much harder to identify successfully (Baldwin and Lui 2010). To show that microblog language is a challenge in itself, we perform an initial experiment on short formal texts versus short microblog texts. In particular, for each language, we use documents from the EuroParl corpus (Koehn 2005) and from those we select sentences <140 characters long. We randomly sample 1,000 sentences per language, from which 500 are used for training and 500 are used for testing. Table 1 shows the performance of our baseline model (detailed in Sect. 3) on the formal (EuroParl) language documents and the microblog posts. Results clearly indicate that language identification on the idiomatic microblog language is more challenging than on formal texts of equal length, with the two systems significantly different according to the *p* test (see Sect. 5 for details on the dataset and significance test).

Table 1 Accuracy for language identification on formal language (EuroParl) and microblog language

	Dutch (%)	English (%)	French (%)	German (%)	Spanish (%)	Overall (%)
Formal	99.6	98.6	99.4	99.4	99.8	99.4
Microblog	90.2	94.8	90.0	95.8	91.2	92.4

To address the effects of very short and ambiguous (in terms of what language) microblog posts, we go beyond language identification on post text alone, and introduce five *semi-supervised priors*. We explore the effects on language identification accuracy of (1) a blogger prior, using previous microblog posts by the same blogger, (2) a link prior, using content from the web page hyperlinks within the post, (3) a mention prior, using the blogger prior of the blogger mentioned in this post, (4) a tag prior, using content of posts tagged with the same tag, and (5) a conversation prior, using content from the previous post in the conversation.

Besides exploring the effects of the individual priors on language identification performance, we also explore different ways of combining priors: we look at post-independent and post-dependent combination models. For the post-dependent combination models, we introduce two ways to measure the confidence of a prior. The confidence of a prior can then be used in a linear combination model. We compare these post-dependent combination models to two post-independent models, (1) a linear combination model with fixed weights, and (2) a voting model.

In particular, we aim at answering the following research questions in this paper: (1) What is the performance of a strong language identification method for microblogs posts? (2) Does domain specific training of language models help improve identification accuracy? (3) What is the effect on accuracy of using priors extracted from microblog characteristics? (4) Can we successfully combine semi-supervised priors in a post-independent way? (5) How can we determine confidence of individual priors? (5) Can we use confidence to combine priors in a post-dependent way?

This paper makes several contributions: (1) it explores the performance of a strong language identification method on microblog posts, (2) it proposes a method to help identification accuracy in sparse and noisy data, (3) it introduces confidence metrics that can be used to weight "sources of evidence", and (4) it performs an in-depth analysis of identification results.

The remainder of the paper is organized as follows: in Sect. 2 we explore previous work in this area. In Sect. 3 we introduce our baseline model, and the semi-supervised priors. Section 4 talks about combining priors, and introduces our confidence metrics. We test our models using the setup detailed in Sect. 5, and in Sect. 6 we present our results. We analyse and discuss the results in Sects. 7 and 8. Finally, we present an analysis of the classification of 1 million tweets published in a single day in March in Sect. 9, and conclude in Sect. 10.

2 Related work

Language identification can be seen as a subproblem in text categorization. Cavnar and Trenkle (1994) propose a character n-gram-based approach to solving text categorization in general, and test it on language identification. Their approach compares a document "profile" to category profiles, and assigns to the document the category with the smallest distance. Profiles are constructed by ranking n-grams in the training set (or the document) based on their frequency. These ranked lists are then compared using a rank-order statistic, resulting in a ‘out-of-place’ (OOP) distance measure between document and category. Tested on a set of Usenet documents, it achieves an accuracy of 99.8 % for language identification.

Other approaches to the n-gram OOP method have been examined in Baldwin and Lui (2010), Bhargava and Kondrak (2010) and Yu et al. (2010). This paper differs in that we examine the utility of microblog priors, as opposed to comparing different classification algorithms. Note that the priors presented in this work could easily be integrated into other models (e.g., Naive Bayes, SVM).

Accuracy is often very high when looking at structured and well-written documents, however research has been done examining different types of text. Language identification on web pages already seems more difficult: Martin and Silva (2005) test an n-gram-based approach with web-related enhancement, and show that accuracy is between 80 and 99 %, depending on the language. Another interesting research by Baldwin and Lui (2010) also explores the impact of document length on language identification. They test language identification on Wikipedia pages, and show that performance on this task improves with growing document length: Accuracy for longer documents reaches 90 %, whereas this is only 60–70 % for shorter documents. Finally, interesting work examining the language identification of query like short text is done by Gottron and Lipka (2010). The authors explore performance of language identification approaches on “queries” (news headlines), which are, on average, 45.1 characters long. They achieve high accuracy results of 99.4 % using 5-grams, but focus on short newswire text, without the idiomatic limitation imposed by the social media domain (the impact of which is demonstrated in Table 1), as examined in this work.

3 Language identification components

Based on previous work, we opt for using an n-gram approach to language identification. More precisely, we use the TextCat[1] implementation of the approach described in Cavnar and Trenkle (1994). This model has shown good and robust performance on language identification. In the previous section we explained how TextCat works to identify a document’s language. We use the TextCat algorithm for language identification on our microblog post set and study the effect on accuracy of language models trained on different data sets. We consider two types of language models: (1) *out-of-the-box*, which uses the training data supplied by TextCat, and we set this as our baseline, and (2) *microblog*, for which we use a training set of posts from our target platform to re-train TextCat.

More formally, let z be the total number of languages for which we have trained language models and $i \in \{1, \ldots, z\}$ denote the corresponding model for a language. For each post p we define a language vector

$$\widehat{\boldsymbol{\lambda}_p} = \langle \lambda_p^1, \lambda_p^2, \ldots, \lambda_p^z \rangle, \tag{1}$$

[1] http://www.let.rug.nl/~vannoord/TextCat/.

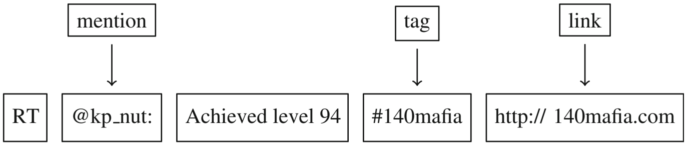

Fig. 1 An example tweet with the three surface features used in our model highlighted

where λ_p^i is a score denoting the distance between post p and language i: the smaller the distance, the more likely it is that post p is written in language i. In the remainder of the paper, we refer to vectors constructed from the microblog post itself as *content-based identification* vectors, written as $\widehat{\boldsymbol{C\lambda_p}}$.

3.1 Semi-supervised priors

On top of the language identification on the content of the actual post, as described above, we use five semi-supervised priors to overcome problems due to sparseness or noise (see Sect. 1) and help improve the accuracy of our baseline classifiers. Our priors are (1) semi-supervised, because they exploit classifications of the supervised language identifier on unlabeled data, for which we do not know beforehand the true language, and (2) priors, because they allow us to identify the language of a post without content-based identification. Given the setting of microblogs, we are offered several natural priors. The example tweet in Fig. 1 shows three surface features we plan to exploit as priors. Besides these three surface features, we also use priors based on the conversation and blogger history.

Link prior: posts in microblogs often contain links, referring to content elsewhere on the web. This content is often of longer text length that the post itself. We identify the language of the linked web page, and use this as link prior for the post that contains the link. Let $L = \{l_1, \ldots, l_k\}$ be a set of links found in post p. For each web page $l_i \in L$ we apply the *out-of-the-box* model to its content, and construct a link prior vector from the average of content-based identification vectors of web pages found in p:

$$ {}_{L}\hat{\boldsymbol{\lambda}}_{\boldsymbol{p}} = \frac{1}{k}\sum_{i=1}^{k} {}_{C}\hat{\boldsymbol{\lambda}}_{l_i}. \tag{2}$$

Blogger prior: behind each post is a blogger who wrote it, and probably the current post is not her first; there is a post history for each blogger the content of which can be beneficial for our purposes. By identifying (or guessing) the language for previous posts by the same blogger, we construct a blogger prior for the current post. Let $P = \{p_1, \ldots, p_k\}$ be a set of posts predating p from blogger u. For each $p_i \in P$, we use the *microblog* language models, and construct $\hat{\boldsymbol{\lambda}}_{\boldsymbol{p_i}}$, as explained before. We then derive a blogger prior from the average of content-based identification vectors of previous posts:

$$ {}_{B}\hat{\boldsymbol{\lambda}}_{\boldsymbol{p}} = \frac{1}{k}\sum_{i=1}^{k} {}_{C}\hat{\boldsymbol{\lambda}}_{\boldsymbol{p}_i}. \tag{3} $$

Mention prior: as a social medium, microblogs are used to communicate directly between people. Post in microblogs are often directed to one or several specific persons, indicated by a special token. We can identify the language for these users that are addressed, and use this information as mention prior. Let $U = \{u_1, \ldots, u_k\}$ be a set of bloggers mentioned in post p. For each $u_i \in U$, we build a blogger prior ${}_{B}\hat{\boldsymbol{\lambda}}_{\boldsymbol{u}_i}$ as in Eq. 3. We derive the mention from the average of blogger priors:

$$ {}_{M}\hat{\boldsymbol{\lambda}}_{\boldsymbol{p}} = \frac{1}{k}\sum_{i=1}^{k} {}_{B}\hat{\boldsymbol{\lambda}}_{\boldsymbol{u}_i}. \tag{4} $$

Conversation prior: certain posts form part of a specific conversation between individuals, as opposed to being part of a more general conversation between numerous bloggers. When this is the case, it is safe to assume that this conversation is taking part in a single language common to both bloggers. Posts that are part of a conversation are not recognizable as such from the content, but this information is stored in the post's metadata. Let p_{i-1} be the previous post in the same conversation as post p. We use the *microblog* language model to construct ${}_{C}\hat{\boldsymbol{\lambda}}_{\boldsymbol{p}_{i-1}}$, as explained before, and use this as the conversation prior ${}_{V}\hat{\boldsymbol{\lambda}}_{\boldsymbol{p}}$.

Tag prior: bloggers often contribute to a corpus of microblog posts on a specific topic, where the topic is represented by a tag. This corpus of posts, i.e., posts that share the same tag, can be beneficial for our purposes. We derive a tag prior based on the average over microblog posts that share the same tag. Let $T = \{t_1, \ldots, t_k\}$ be a set of posts predating p in the corpus of tag T. For each $t_i \in T$, we use the *microblog* language models, and construct ${}_{C}\hat{\boldsymbol{\lambda}}_{\boldsymbol{p}_i}$, as explained before.

$$ {}_{T}\hat{\boldsymbol{\lambda}}_{\boldsymbol{p}} = \frac{1}{k}\sum_{i=1}^{k} {}_{C}\hat{\boldsymbol{\lambda}}_{\boldsymbol{t}_i}. \tag{5} $$

Since scores generated by TextCat are not normalized by default, for all priors that require averaging, that is all those except the conversation prior, we normalize the raw scores using z scores. Our language identification approach leaves us with a content-based identification vector and five semi-supervised priors. For ease of reference, in the rest of the paper, priors will refer to these five priors and the content-based identification vector, unless clearly stated otherwise. The next section details how we combine these vectors into one, and obtain our final estimate of a tweet's language.

4 Combining priors

The combination of priors and the content-based identification is a kind of "evidence combination" and we have two obvious ways of going about it: (1) treat all posts equally, and use post-independent combination models, or (2) observe each post individually, and use a post-dependent model to combine evidence. For the first combination approach we need training data, and for the second approach we need a way to determine which priors are most reliable for a given post. In this section we explore both aspects: Sect. 4.1 introduces the post-independent combination models and Sect. 4.2 discusses the post-dependent combination, with a focus on the *confidence metrics* that can be used. After discussing our models and metrics here, we introduce our dataset in the next section and discuss how we train our models.

4.1 Post-independent combination

In this section we present two different ways for post-independent prior combination. The first approach uses post-independent weight optimization for linear interpolation and the second is based on voting, a technique for combining multiple classifiers.

4.1.1 Linear interpolation with post-independent weight optimization

To create a linear model, we first construct vectors for the content, and each of the priors, with scores for each language, and combine these vectors using a weighted linear combination. More formally, we identify the most probable language for post p as follows:

$$lang(p) = argmax \sum^{q} w_q \cdot {}_{\boldsymbol{q}}\hat{\boldsymbol{\lambda}}_{\boldsymbol{p}}, \tag{6}$$

where $q = \{L, B, M, C, V, T\}$. This model has two important components: first, to make ${}_{\boldsymbol{q}}\hat{\boldsymbol{\lambda}}_{\boldsymbol{p}}$ suitable for linear interpolation, we need to normalize the values. Scores are normalised using z scores. The second component is w_q, the actual weight of the prior q. To find the optimal weights for each prior, we perform a sweep over the parameter space in an interpolated model over all priors. We optimize for overall accuracy (accuracy over all five languages) on our development set (see Sect. 5). The post-independent weight optimization approach does not take post-specific features into account and requires training data for the weights.

4.1.2 Majority voting

As well as trying sweeps for the optimal linear interpolation parameters, we explore the use of voting for classifying a post. Majority voting is a principled way to combine classifications from multiple classifiers (Dietterich 2000). Majority voting applies each classifier to the input, in this case a post, takes the classifications, and selects the label that was assigned most. As long as each individual classifier performs better than chance, it has been shown that this approach can lead to a better performance than relying on any single classifier (Dietterich 2000).

The main issue with majority voting is how to deal with ties: the case where multiple labels receive an equal number of votes. In our case, we use the normalized scores for solving ties. When a tie occurs, we select the label (language) that has the highest normalized score over all priors. Although more ways of solving ties are possible, experiments on the development set show this approach is successful. The advantage of the majority voting approach is that it is quite insensitive to fluctuations in scores, since it only relies on votes. On the other hand, ignoring scores also means the loss of (potentially valuable) information on the strength of priors.

4.2 Post-dependent combination

The aim of a post-dependent model is to vary the weights of the priors that give optimal classification results for that specific post. Here, we propose to use a post-dependent linear combination model. This model is similar to the one introduced in Eq. 6, where each prior is weighted. Unlike the post-independent linear interpolation, however, we cannot learn these weights, since we only have one instance from each post. In this section, we introduce two ways of estimating the *confidence* of each prior, which can be used in our linear combination.

To explain the notion of confidence, observe the two situations in Fig. 2. The top half shows a situation where the prior is very confident: one language (the black dot) is close to the post (white dot), and the other languages (shaded dots) are quite far away. This prior is confident that this post is written in a certain language. The bottom example shows a different situation, in which several languages (shaded dots) are close to the post: the prior is uncertain as to which language is the right one. We aim to exploit the observations from Fig. 2, and propose the following two confidence metrics: (1) the beam confidence, and (2) the lead confidence.

4.2.1 Beam confidence

The beam confidence builds on the observation that when multiple languages are close to the most likely language, the prior is less confident. To concretize this observation, we use the following reasoning: Given a beam b (e.g., 5 %), we calculate a limit distance based on the (raw) distance of the most likely language.

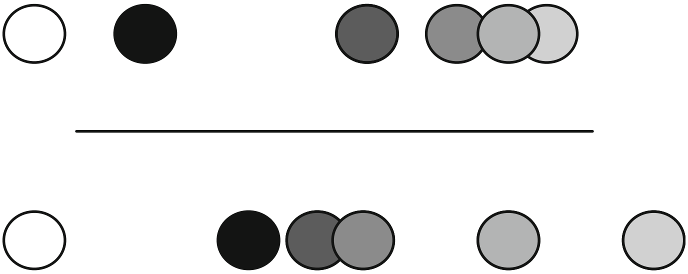

Fig. 2 Two graphical representations of confidence, with a confident prior (*top*) and uncertain prior (*bottom*). The *white dot* represents the post profile and the *shaded dots* represent the profiles of different languages

Languages are ordered by their raw scores, from lowest to highest. The language first in this list is the most likely. This limit distance is defined as *limit*(p) = $d(\lambda^1) + b$, the raw distance of the most likely language increased by the beam (in percentages). We then move on to the next most likely language, and see if this language is closer to the post profile than the limit. If this is the case, we add this language to the list of languages "within the beam", *LIB*(p), and repeat with the next most likely language. If not we stop. Eqs. 7 and 8 show how we calculate the *LIB*(p) for post p over all languages λ.

$$LIB(p) = \sum_{i=2}^{k} inBeam(\lambda^i) \tag{7}$$

$$inBeam(\lambda^i) = \begin{cases} 1 & \text{if } d(\lambda^i) < d(\lambda^{i-1}) + b \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $d(\lambda^i)$ is the raw distance between the post profile and the language profile.

We now have the number of languages that falls within the beam; from this we can calculate a weight for the prior. We use both a linear and exponential function to obtain the final weights. The linear function is defined as follows:

$$weight(p) = \frac{|\lambda| - LIB(p)}{|\lambda| - 1} \tag{9}$$

The exponential function uses an exponentially increasing punishment for more language in the beam:

$$weight(p) = e^{-LIB(p)+1} \tag{10}$$

4.2.2 *Lead confidence*

The second confidence metric we introduce is the lead confidence. This metric tries to capture the lead the most likely language has over its closest "rival" language. The further away a language is from its nearest rival, the more confident a prior apparently is about this language. We use a fairly straightforward approach to measure the lead confidence: we take the difference between the first $d(\lambda^1)$ and second $d(\lambda^2)$ ranked languages normalized scores. We take this difference as the weight of the prior:

$$weight(p) = d(\lambda^1) - d(\lambda^2) \tag{11}$$

5 Experimental setup

For testing our models we need a collection of microblog posts. We collect these posts from one particular microblog platform, Twitter. We test our models on a set of five languages, Dutch, English, French, German, and Spanish, and gather an initial set of *tweets* (Twitter posts) by selecting tweets based on their location. From this initial sample, we manually select 1,000 tweets in the appropriate language:

Table 2 Number of tweets in the training set (400 tweets per language) with at least one link, tag, or reply and the total number of these items per language

Language	Number of tweets with			Total number of		
	Links	Tags	Replies	Links	Tags	Replies
Dutch	59	77	213	60	94	251
English	123	54	174	123	78	201
French	140	71	183	143	105	217
German	182	107	108	183	219	119
Spanish	103	42	190	103	55	226

tweets that contain only non-word characters (e.g. URLs, emoticons) are ignored. For multilingual tweets, we assign the language that is most "content-bearing" for that post.

For training purposes, we split each set in a training set of 400 tweets (for TextCat training), a development set of 100 tweets (for weight optimization), and a test set of 500 tweets.[2] For the blogger prior, we extract as many tweets as possible from the poster's history, which on average is 154.8 posts per user. For the mention prior, of the 2,483 unique users mentioned in tweets, the average number of tweets extracted from the posters history was 129.3. For the hashtag prior, we extract the 200 most recent posts that contain the hashtag posts. We placed no time restrictions on the extraction of such data. Table 2 lists several characteristics of the tweets in our training set.

TextCat allows us to select the number of n-grams we want to use for profiling our language and documents. Preliminary experimentation with this parameter revealed that the standard value (top 400 n-grams) works best, and we use this value for the remainder of the experiments. We report on accuracy (the percentage of tweets for which the language is identified correctly) for each language, and overall.

The number of languages examined will impact on the absolute accuracy results reported, both for the baseline system and for the more elaborate methods proposed here. However, our goal in answering the six research questions is to demonstrate a significant increase in performance over the baseline using the methods proposed in this work. For computing significance between two models, we use the p test (Yang and Liu 1999) on the overall accuracy:

$$Z = \frac{p_a - p_b}{\sqrt{2p(1-p)/n}},$$

where $p = \frac{p_a + p_b}{2}$, p_a and p_b are accuracy results of the two systems being compared, and n is the number of tweets classified by both models. Significance levels of 90, 95 and 99 % are referred with $^{!}$, †, and ‡ respectively.

[2] The training data and the trained models are available at http://ilps.science.uva.nl/resources/twitterlid.

6 Results

We design and conduct four experiments to answer our six research questions. Below, we detail each of the four experiments and present the results.

Language identification on microblog posts The first experiment aims at answering the first two research questions, namely, what is the performance of a strong language identification method on microblog posts, and whether domain specific training can help improve accuracy. Results in Table 3 show that language identification on short posts in microblogs is not as straightforward as it is on formal short pieces of text (see Table 1, where accuracy on formal text is much higher). The use of the microblog model improves performance by 3 % on average, but accuracy is still limited, with Dutch showing no improvement at all.

Individual priors In our second experiment we target our third research question and we study the effect on accuracy of our set of individual semi-supervised priors which we derived from microblog characteristics. We learn the weights of the prior versus the content-based identification on our development set using weight sweeps as explained in Sect. 4.1.1, limiting the sum of weights to 1, and report on the best performing prior weights in Table 4. The results show that incorporating the semi-supervised priors leads to an increase in accuracy for all languages over content-based identification using the microblog model. In particular, among all priors, the blogger and mention priors are found to perform the best, as they encode the language in which the blogger usually posts, and the language of the blogger's social network.

Post-independent In our third experiment we tackle research question four. Here, we look at the effect on performance after we combine individual priors in a post-independent way. We learn the weights as explained before and find that the content-based identification vector (0.4), blogger prior (0.3), link prior (0.1), and mention prior (0.2) contribute to the best performing setting. Table 5 (top) shows that combining the priors results in better accuracy than using them individually. In particular, performance peaks when we make use of fixed weights in the linear interpolation. Inspection of the results reveals that most errors in the voting method are due to ties, which, according to the results, are not always handled appropriately by our method.

Post-dependent In our last experiment, we turn to our last two research questions, namely, the effect of post-dependent combination of priors and the use of different confidence scores of priors. Before testing, we explore the beam function and width for the beam confidence. Experiments on the development set show a clear preference for the exponential function (95.4 vs. 91.0 % accuracy using a 10 % beam). As to the beam width *b*, we look at values of 1, 5, 10, and 15 % using the

Table 3 Results for baseline content-based identification runs using the out-of-the-box and the microblog language models

	Dutch (%)	English (%)	French (%)	German (%)	Spanish (%)	Overall (%)
Out-of-the-box	**90.2**	88.4	86.2	94.6	88.0	89.5
Microblog	**90.2**	**94.8**	**90.0**	**95.8**	**91.2**	**92.4**$^!$

For each language, the model with the highest accuracy has its score in bold

Table 4 Results for content-based identification and five individual semi-supervised priors using the microblog language model

Run	Dutch (%)	English (%)	French (%)	German (%)	Spanish (%)	Overall (%)
Microblog	90.2	94.8	90.0	95.8	91.2	92.4
Blogger (0.4)	**95.2**	**98.6**	**95.4**	**98.6**	**96.0**	**96.8**‡
Link (0.2)	90.2	95.4	90.6	96.2	91.8	92.8
Mention (0.3)	91.6	96.0	90.8	96.6	93.0	93.6
Tag (0.2)	90.4	95.2	90.4	96.0	91.4	92.7
Conv. (0.3)	90.8	95.4	90.6	96.2	92.2	93.0

For each language, the model with the highest accuracy has its score in bold

The weights assigned to each prior are shown in brackets, and learnt on the development set. We test for significant differences against the baseline microblog model

Table 5 Results for content-based identification runs using *post-independent* (§4.1; lines 3 and 4) and *post-dependent* (§4.2; lines 5 and 6) combination of the priors and the microblog language model

Run	Dutch (%)	English (%)	French (%)	German (%)	Spanish (%)	Overall (%)
Blogger (0.4)	95.2	98.6	**95.4**	98.6	96.0	96.8
Linear int.	96.0	99.0	**95.4**	**98.8**	**96.8**	97.2‡
Majority vote	94.4	96.4	94.2	97.2	**96.8**	95.8†
Beam conf.	**97.6**	**99.4**	95.2	98.6	96.2	**97.4**‡
Lead conf.	96.0	99.2	90.6	97.8	94.4	95.6†

For each language, the model with the highest accuracy has its score in bold

We test for significant differences against the microblog + blogger model

exponential function. Here, the difference is not as big, but we find that 5 % is most favorable (97.8 vs. 97.6 % for 1 % beam and 95.4 % for 10 % beam). Results in Table 5 (bottom) show that post-dependent combination outperforms the use of individual priors and is marginally better than post-independent combinations.

Turning to accuracy for individual languages, we see that language identification works best for English and German, followed by Dutch, French and Spanish with performance hovering at the same levels. In the next section we briefly touch on this with some examples of errors made in the identification process.

7 Error analysis

In analyzing the posts misclassified by our final classifier using all priors, we group them into four distinct categories: fluent multilingual posts, those containing named entities, prior effects, and language ambiguous. We give examples in Table 6, and explain each type of error in turn.

Fluent multilingual posts: these are posts which are a grammatical sentence with words written in two or more languages. Usually these take the form of a sentence split into two, with both halves in different languages.

Table 6 Examples of misclassified tweets, along with the languages assigned, broken down by error type

Language		Content of microblog post
Assessed	Classified	
Fluent multilingual posts		
French	English	RT @msolveig: Sinusite de printemps, pause pour le moment… V.I.P. reporté, qqs jours de repos et je serai sur pieds. Sorry… Good luck!!!
Spanish	English	RT @FlamencoExport: Espana no solo es flamenco. Tambien es jamon! RT @Plateofjamon Nice article about Iberian ham: http://nyti.ms/6QVF9I …
Posts containing named entities		
French	English	Vous insultez Ashley de pouf ,de pétasse et autre … mais vous vous êtes vu bande de connasse ? #JeMenerve
Spanish	English	Pues yo slo quiero que venga Panic! At The Disco. Con eso me conformo.
Prior effects		
French	English	EPISODE N^{o} 2 : DANS LA LAGUNE…: http://bit.ly/bhi4FG #buzz
Spanish	English	@mariaam1004 *-* Graciaaas! Mi tweet #4777 va para tí (:
Language ambiguous posts		
French	English	#emploi #technicien TECHNICIEN(NE) BE ELECTRIQUE http://is.gd/bnx8A
Dutch	English	@Chenny83 Ja :D

Named entity errors: these posts are misclassified because they contain a reference to a foreign language named entity, such as a company or product name, song title, etc. The named entities contained in the post outweigh the correct language tokens in the post in scoring, leading to the misclassification.

Prior effects: the use of priors can sometimes have a negative effect. For example, if the user mentioned a post in a different language to their own post, or when a tag is used mostly from a different language group. E.g., some tweets contain links which point to a webpage in a different language to that used in the post.

Language ambiguous: these posts are misclassified because they only contain a few tokens which could belong to a number of different languages.

Finally, we demonstrate in Table 7 for each true language the number of tweets which were incorrectly assigned another language for the post-dependent beam microblog model. In the final row we show the total counts for each misclassified language. English is the most incorrectly assigned label by far, with 54 out of 65, or 83 %, of misclassified tweets being assigned an English label. French, as demonstrated in Table 5, has the most misclassified posts.

8 Discussion

We discuss how the weights of individual priors affect performance, the robustness of our methods when domain specific training is unavailable, and finally candidate priors unexplored in this paper for methodological reasons.

Table 7 Misclassification breakdown by language

	Dutch	English	French	German	Spanish	Total
Spanish	1	17	0	1	–	19
German	0	7	0	–	0	7
French	1	21	–	0	2	24
English	1	–	0	0	2	3
Dutch	–	9	1	2	0	12
Total	3	54	1	3	4	65

The leftmost column represents the correct language, and numbers indicate the number of posts classified as another language. Finally in the rightmost column we show the total number of misclassified posts per language

8.1 Individual prior weights

In order to better understand the effects of individual priors when combined with the content-based identifier, we vary their weights from not using the prior at all (0), to using almost only the prior and not the content (0.9). Figure 3 shows that for prior weights around 0.4 priors are most helpful. Blogger, mention, and conversation priors are robust to the weights, whilst link and tag show a drop in performance when they are weighted more than 0.4.

8.2 Domain nonspecific training

As shown earlier in Table 3, training on microblog posts clearly outperforms the use of out-of-the-box models supplied with TextCat. However it may not always be possible to acquire the microblog posts for training, especially if applying the language identifier to many languages. To examine the improvements possible when using out-of-the-box models (or data from domains other than microblogs), we show in Table 8 results using priors trained on these models.

The best results using a single prior are achieved using the blogger prior, giving 5 % improvement in overall classification accuracy over a domain generic baseline. Again, the combinations of priors show best overall accuracy, with the linear interpolation (post-independent) and the beam confidence (post-dependent) resulting in a 6.5 % increase. Interestingly, the best reported accuracies using out-of-the-box models are only about 1.5 % lower than best reported microblog models, indicating that, if it is not possible to acquire microblog posts for training, using normal text with the priors defined in this paper can still lead to high classification results.

9 Online twitter analysis

Usage of Twitter is not just limited to the English-speaking world. Other countries, like Indonesia, Brazil, Germany, and the Netherlands actively participate on

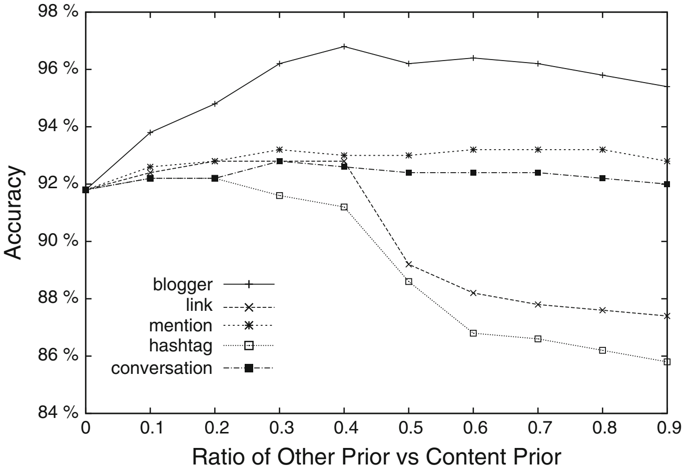

Fig. 3 Accuracy while changing the ratio of individual priors and content-based prior

Twitter, and contribute to a large degree to what is discussed in the microblogosphere. However the distributional profiles of language use on Twitter remains unknown, and thus, alongside the work published in Hong et al. (2011), Poblete et al. (2011) and Semiocast (2010) we provide the one of the first analyses of language use on Twitter.

9.1 Twitter language distribution

We apply our language identification method to a corpus of 1.1 million tweets, collected during the period of 1 day (2nd of March 2011). These tweets are collected from the GardenHose Streaming API service provided by Twitter, which

Table 8 Results and significance levels for content-based identification, five individual semi-supervised priors, and their combinations using the TextCat language model: *blogger, link, mention, tag, conversation*

Run	Dutch (%)	English (%)	French (%)	German (%)	Spanish (%)	Overall (%)
Out-of-the-box	90.2	88.4	86.2	94.6	88.0	89.5
Blogger (0.4)	95.6	95.8	91.4	98.6	92.0	94.7‡
Link (0.2)	90.0	88.8	86.4	95.0	87.4	89.5
Mention (0.3)	92.0	90.6	87.0	95.0	89.8	90.9
Tag (0.2)	90.2	89.0	85.6	95.0	87.8	89.5
Conv. (0.3)	91.4	89.0	86.6	95.0	89.2	90.2
Linear int.	96.4	96.6	**91.8**	**98.8**	93.2	95.4‡
Majority vote	95.0	**98.0**	89.2	97.4	93.2	94.6‡
Beam conf.	**97.0**	97.8	**91.8**	98.2	**94.8**	**95.9**‡
Lead conf.	94.0	97.8	86.0	96.6	90.8	93.0†

For each language, the model with the highest accuracy has its score in bold

Table 9 Number of tweets with at least one link, tag, or reply and the total number of these items in the set of 1.1 million tweets

	Link	Tag	Reply
Number of tweets	204,127	141,457	621,122
Total number	205,624	191,625	819,553

represents a random sample of 10 % of the public posts on Twitter. For the languages that fall outside of our original five languages, we use the language models distributed with TextCat. In Table 9 we provide the feature statistics of this corpus over all languages.

In Fig. 4, we present the ranked distribution of post languages with counts over 1,000. English ranks highest, with Japanese and Spanish following in second and third. Together, they make up approximately 63 % of corpus. The top five languages make up 82 % of all tweets in our corpus, and the top 10 languages make up 92 %.

The presence of Esperanto and Latin posts is surprising. A manual evaluation confirms these can be accounted for due to classification error.[3] The approximately 1,000 tweets classified as Latin and Esperanto represent only a small portion of the entire corpus (0.009 %). The findings published in other work (Hong et al. 2011; Poblete et al. 2011; Semiocast 2010) independently confirm the validity of the reported results in this work with respect to the top languages used on the Twitter microblogging platform.

Having a large corpus of labeled microblog posts, we now turn our attention to answering the following analysis questions:

1. Does language use alter with time of day?
2 To what extent do the classified languages correlate with the geo-location and language metadata fields that come with the Twitter stream?
3. How does usage of Twitter features (used as priors in this work) change with language?

9.2 Time series analysis

Examining the corpus of 1.1 million tweets, we do not know the true underlying distribution. A manual evaluation of all 69 languages classified in the corpus is not possible by the authors. However, we believe it would be interesting to examine the language use of bloggers with time. In particular, we expect to see differences in language use of the top five languages classified according to different time zones. Using the 'created_at' time field within the metadata, we bucket each post by their publication hour. Hours are based on GMT +0000. We present the results in Fig. 5.

We can clearly see two groups of languages according to their distribution over time: (1) English, Spanish, and Portuguese and (2) Japanese and Indonesian. The former group of languages has its largest speaking population in the Americas,

[3] Note we do not claim that our language identification classification system achieves 100 % accuracy, and thus the inclusion or absence of certain languages could be a result of incorrect labeling.

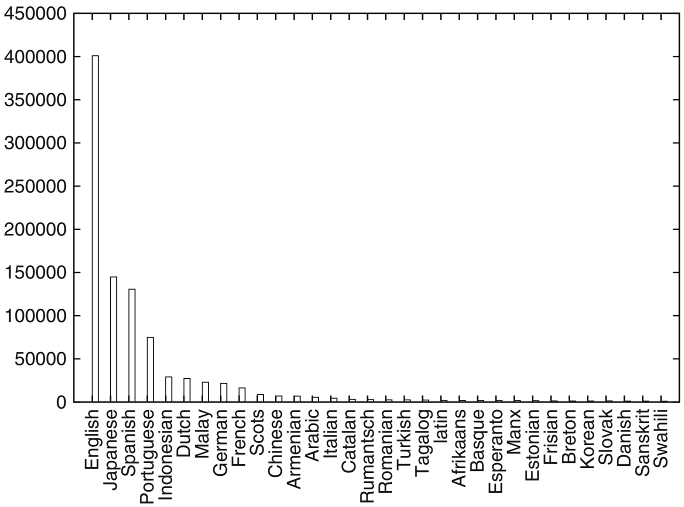

Fig. 4 Number of tweets per language (published during 1 day, showing languages with >1,000 posts)

including the USA (English), Brazil (Portuguese), and the other South American countries (Spanish). The latter group is mainly focused around Japan and Indonesia. The differences in time zones between the countries in the two groups explain the differences in peak and dip times: The Asian languages peak around 1 p.m. GMT and reach their lowest dips around 8 p.m. GTM. For the other group of languages we find the peaks between 11 p.m. and 3 a.m., and their dips are found at 7–9 a.m. GMT.

Converting the GMT times to the actual times of the main contributing countries for each language group, we find that for both group the peaks appear between 10 p.m. and midnight and the dips are in the early morning (4–5 a.m.).

9.3 Metadata

Obvious priors to explore when creating a language identifier for microblog platforms are the metadata fields that could hint towards the language used by the blogger. Twitter offers two such fields, geo-location and interface language. The geo-location prior was left unexplored in Sect. 6 for methodological reasons: In order to collect tweets for a language for our experiments, we used the location to filter tweets for that language. Using location as a prior would be biasing the results. We also ignored the interface language field, as it is limited seven languages (English, French, German, Italian, Japanese, Korean, Spanish). Having classified a large corpus of tweets, it is interesting, though, to see to what extent these metadata fields correlate with the languages assigned.

Interface language field. In Table 10 we present the distribution of languages according to the interface language metadata field, along with the number of tweets assigned to each of the seven languages according to our own classifier. Interestingly, of the seven language options offered by Twitter, our classifier and

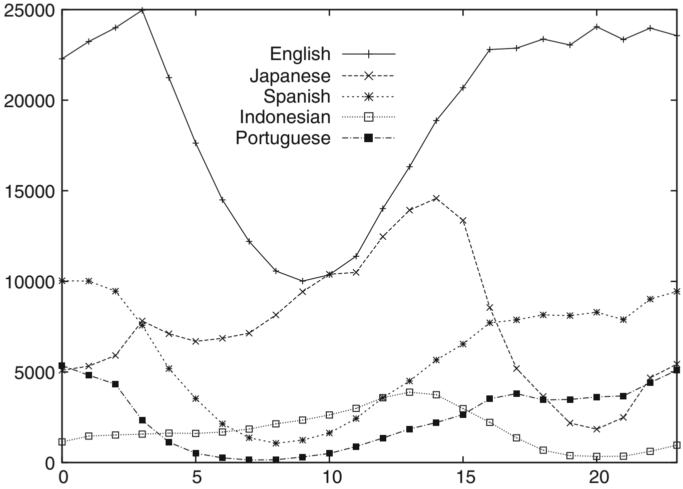

Fig. 5 Number of tweets in each language published on Twitter in hourly buckets (hours 0−23), for the top five languages classified

the language metadata field only agree on Japanese, with a difference of 5,000 tweets. For English, we find that almost two times more tweets come from the English interface than our classifier assigns to English (840,000 vs. 460,000). We observe similar patterns for German and Korean, while the effect is reversed for French, Spanish, and (less so) for Italian. These figures, along with the fact that there are many more languages used in microblog posts than are offered as interface language options (including Portuguese and Indonesian), indicate the poor suitability of the interface language field in itself as a predictor of microblog post language.

Geo-location field. We now turn our attention to the analysis of the geo-location information. In particular, Twitter automatically encodes the longitude and latitude points into country information. In total, only 17,737 of the 1.1 million tweets contain geo-location information, with 34 countries presented in total. The top countries according to this field are Brazil (6,527 tweets), USA (4,616), Indonesia (2,080), the UK (1,164), and the Netherlands (500). Due to the sparsity in use of the geo-location information, we posit the utility of the geo-location field for language identification as a prior within our framework is limited.

9.4 Twitter feature usage

We are interested in the way people use Twitter in different languages, and would like to see if there are obvious differences between languages in the usage of Twitter features. For this, we look at three Twitter specific features, *hashtags, links* and *mentions*, and explore their usage in the top five languages classified.

In Table 11 we report on the percentage of tweets that contain a link for each language, the percentage of tweets having at least one hashtag, the average number

Table 10 Tweets per language according to the language metadata field and our classifier

	English	French	German	Italian	Japanese	Korean	Spanish
Metadata	839,856	8,150	6,450	3,348	185,360	6,657	101,728
Classified	459,318	42,706	4,890	4,890	180,140	1,077	142,401

of hashtags per *tagged* tweet, the percentage of tweets that contain at least one mention and finally the average number of mentions in tweets that have mention.

We see that Indonesian and Spanish show high mention usage, with over three quarters of Indonesian tweets containing at least one mention. On average, they contain 1.8 mentions, indicating the popularity of this feature for Indonesian microbloggers to interact with multiple other microbloggers. We also find that for the other languages the popularity of mentions does not influence the number of mentions per tweet.

The proportion of tweets containing a tag or link is far lower across all languages than those containing mentions. English and Spanish have the highest percentage of tweets containing a hashtag. Though only 10.8 % of Portuguese tweets contain a hashtag, when they do, they have the highest average tags per tagged tweet rate, indicating that when they do use tags, they tend to use multiple. Finally, English displays the highest proportional use of links, with just over 25 % containing a link, 10 % more than Spanish posts at 14.4 %.

10 Conclusion

In this paper we study language identification on microblog posts. We have demonstrated that, given the short nature of the posts, the rather idiomatic language in these (due to abbreviations, spelling variants, etc.), and mixed language usage, language identification is a difficult task.

Our approach is based on a character n-gram distance metric. To tackle the challenges in microblogs, we identify five microblog characteristics that can help in language identification: the language profile of the blogger (blogger), the content of an attached hyperlink (link), the language profile of other users mentioned (mention) in the post, the language profile of a tag (tag), and the language of the original post (conversation), if the post we examine is a reply. Further, we look at methods on how to combine these priors in a *post-dependent* and *post-independent* way.

Results show that the use of language models trained on microblog posts increase accuracy by 3 %. Individual priors add to performance, with the blogger prior adding another 5 %. The combination of priors is found to outperform their individual use, with post-dependent combination leading to the best performance, close to that of formal short texts. A manual analysis reveals four main categories of errors: fluent multilingual posts, prior effects, named entity errors, and language ambiguity.

Table 11 Twitter feature usage per language, for the top five languages

	English	Japanese	Spanish	Portuguese	Indonesian
Mentioned tweets	54.8 %	48.2 %	61.6 %	44.8 %	76.6 %
Avg. mentions per tweet	1.3	1.2	1.3	1.2	1.8
Tagged tweets	16.6 %	4.17 %	13.8 %	10.8 %	9.8 %
Avg. tags per tweet	1.4	1.2	1.2	1.5	1.1
Linked tweets	26.1 %	10.7 %	14.4 %	10.0 %	12.4 %

We also conducted a large-scale study of language distribution on a popular, global, microblogging platform. We have demonstrated that the language and country metadata fields that come with the microblog posts make poor signals for language identification, with the language field greatly over-or-underestimating the true underlying language distribution, and the geo-location field being too sparsely used to be relied upon for language identification. Finally, we have demonstrated the differing use of Twitter specific features per language.

We leave to future work the resolution of multi-coded tweets, including the construction of more complex models that are sensitive to within-post language change, possibly via latent methods. Further, the explicit handling of an ‘other’ or ‘unknown’ category would prove beneficial for real-world systems, and more sophisticated approaches to combining priors, such as data fusion, may be worth investigating. Finally, although most priors examined in this work are specific to microblogging, certain features could be tested with respect to a Short Message Service (SMS) corpus.

Acknowledgments We would like to thank the reviewers for their constructive comments and valuable suggestions. This work was supported by the MataHari Project, by the Netherlands Organisation for Scientific Research (NWO) under project numbers 612.061.815 and 380.70.011, and by the European Union’s ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement number 250430 (GALATEAS).

References

Altheide, D. L. (1996). *Qualitative media analysis (Qualitative research methods)*. London: Sage Pubn Inc.

Baldwin, T., & Lui, M. (2010). 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* (*NAACL HLT 2010*) (pp. 229–237). Association for Computational Linguistics.

Bhargava, A., & Kondrak, G. (2010). Language identification of names with SVMs. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (*HLT-NAACL 2010*) (pp. 693–696). Association for Computational Linguistics.

Carter, S., Tsagkias, M., & Weerkamp, W. (2011). Semi-supervised priors for microblog language identification. In *Dutch-Belgian information retrieval workshop* (*DIR 2011*).

Cavnar, W., & Trenkle, J. (1994). N-gram-based text categorization. In *Proceedings of third annual symposium on document analysis and information retrieval* (pp. 161–175).

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Proceedings of the first international workshop on multiple classifier systems* (pp. 1–15).

Golovchinsky, G., & Efron, M. (2010). Making sense of twitter search. In *Proceedings of CHI 2010 workshop on microblogging: What and how can we learn from it?*.

Gottron, T., & Lipka, N. (2010). A comparison of language identification approaches on short, query-style texts. In *Advances in Information retrieval, 32nd European conference on IR research* (*ECIR 2010*) (pp. 611–614).

Hong, L., Convertino, G., & Chi, E. (2011). Language matters in twitter: A large scale study. In *Proceedings of AAAI conference on weblogs and social media* (*ICWSM 2011*) (pp. 518–521).

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology, 60*(11), 2169–2188.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In: MT Summit.

Martins, B., & Silva, M. (2005). Language identification in web pages. In *Proceedings of the 2005 ACM symposium on applied computing* (pp. 764–768).

Massoudi, K., Tsagkias, M., de Rijke, M., & Weerkamp, W. (2011). Incorporating query expansion and quality indicators in searching microblog posts. In *33rd European conference on information retrieval (ECIR 2011)* (pp. 362–367). New York: Springer.

Oghina, A., Breuss, M., Tsagkias, M., & de Rijke, M. (2012). Predicting IMDb movie ratings using social media. In *34th European conference on information retrieval* (*ECIR 2012*). New York: Springer.

Poblete, B., Garcia, R., Mendoza, M., & Jaimes, A. (2011). Do all birds tweet the same? Characterizing twitter around the world. In *Proceedings of the 20th ACM international conference on information and knowledge management* (*CIKM 2011*) (pp. 1025–1030).

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web* (*WWW 2010*) (pp. 851–860).

Semiocast. (2010). Half of messages on twitter are not in English, Japanese is the second most used language. http://semiocast.com/downloads/Semiocast_Half_of_messages_on_Twitter_are_not_in_English_20100224.pdf.

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In International AAAI Conference on Weblogs and Social Media (ICWSM 2010), (pp. 178–185).

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on human factors in computing systems* (*CHI 2010*) (pp. 1079–1088).

Weerkamp, W., Tsagkias, M., & Carter, S. (2011). How people use twitter in different languages. In *Proceedings of the 3rd international conference on web science* (*WebSci 2011*).

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 42–49).

Yu, D., Wang, S., Karam, Z., & Deng, L. (2010). Language recognition using deep-structured conditional random fields. In *2010 IEEE International conference on acoustics speech and signal processing* (*ICASSP 2010*) (pp. 5030–5033).