Soc. Netw. Anal. Min. (2015)5:35
DOI 10.1007/s13278-015-0275-z

ORIGINAL ARTICLE

Bridging social media via distant supervision

Walid Magdy[1] · **Hassan Sajjad**[1] · **Tarek El-Ganainy**[1] · **Fabrizio Sebastiani**[1]

Received: 10 March 2015 / Revised: 9 June 2015 / Accepted: 16 June 2015

Abstract Microblog classification has received a lot of attention in recent years. Different classification tasks have been investigated, most of them focusing on classifying microblogs into a small number of classes (five or less) using a training set of manually annotated tweets. Unfortunately, labelling data is tedious and expensive, and finding tweets that cover all the classes of interest is not always straightforward, especially when some of the classes do not frequently arise in practice. In this paper, we study an approach to tweet classification based on distant supervision, whereby we automatically transfer labels from one social medium to another for a single-label multi-class classification task. In particular, we apply YouTube video classes to tweets linking to these videos. This provides for free a virtually unlimited number of labelled instances that can be used as training data. The classification experiments we have run show that training a tweet classifier via these automatically labelled data achieves substantially better performance than training the same classifier with a limited amount of manually labelled data; this is advantageous, given that the automatically labelled data come at no cost. Further investigation of our approach shows its robustness when applied with different numbers of classes and across different languages.

Keywords Twitter · YouTube · Tweet classification · Distant supervision

1 Introduction

Interest in classifying microblogs has increased with the widespread use of microblogging platforms such as Twitter. Tweets contain useful information that can be applied to various tasks, such as mass emergency management (Imran et al. 2014), stock market analysis (Bollen et al. 2011), social studies (Dodds et al. 2011), and many others. Classifying tweets is usually an essential step in most such applications. From time to time, this may take the form of classification by topic, by sentiment, by political leaning, etc. One of the classification tasks that has received some (although still insufficient) attention is classifying tweets into general-purpose classes, such as e.g. Politics, Sports, Entertainment, Science, etc. Pre-classifying tweets under general-purpose classes can be useful in many applications, such as in online market research and advertising, social analysis of groups' or individuals' interests, and social search. In general, classifying tweets under general-purpose classes is an important enabling technology for applications that attempt to make sense of the Twitter firehose.

Classifying tweets involves several challenges. First of all, tweets contain a variety of information on a variety of topics, and given a specific tweet it is not easy to define an exact class for it. Consider the tweet

Fabrizio Sebastiani is on leave from Consiglio Nazionale delle Ricerche, Italy.

✉ Walid Magdy
wmagdy@qf.org.qa

Hassan Sajjad
hsajjad@qf.org.qa

Tarek El-Ganainy
telganainy@qf.org.qa

Fabrizio Sebastiani
fsebastiani@qf.org.qa

[1] Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

Published online: 04 July 2015

The funniest reaction of a Barcelona supporter after the great goal by Messi `youtu.be/jLeTMIoAgCw`

This tweet could be classified into classes such as Sports, Comedy, or Entertainment. As far as we know, there is no standard set of classes defined for microblogs that accommodate the variety of information available on Twitter. Most microblog classes defined in previous works were motivated by specific applications (Becker et al. 2011; Chen et al. 2011; De Choudhury 2012; Irani et al. 2010; Kothari et al. 2013; Sankaranarayanan et al. 2009), and the number of classes was usually limited to a small number, typically five or less. The only work we are aware of that uses a substantively larger number (18) of microblog classes is Lee et al. (2011); however, in this work, the classes were derived from tweeting trends popular during a certain period ("trending topics"), and these trends tend to change over time.

A major challenge for standard classification approaches is the fact that manually annotated data are required to train an effective classifier. Data annotation is an expensive and time-consuming task, especially when a large number of classes is used, since a sufficient number of examples per class are required to yield reasonable classification accuracy. Sometimes, finding tweets that cover all the classes of interest is not straightforward, especially for classes that do not frequently arise in practice.

In this paper we present a novel method, based on distant supervision, for automatically deriving standard class labels for tweets, so as to generate a large number of training examples for microblog classification. Our proposed method does not require any manual annotation. We use crowdsourced labels from another social medium, YouTube, and we use these labels for training a single-label multi-class tweet classifier (i.e., a classifier entrusted with assigning exactly one class out of several possible ones to each tweet). The benefits of using this method stem from the practically unlimited availability of training instances of this type. Furthermore, this method is language independent; this makes it easy to train a classifier for tweet classification for any language available on Twitter.

We have collected a large set of English-language tweets linking to YouTube videos. Each YouTube video is assigned to a class out of 18 predefined classes as a requirement when posting the video. We apply the class assigned to a YouTube video to the tweets which link it, which creates a large set of automatically labelled microblog data. We have then trained a classifier using hundreds of thousands of tweets linking to YouTube and covering 14 classes; in the literature, this is usually called *distant supervision* (Go et al. 2009). We have then used this classifier to classify unlabelled tweets (not necessarily containing links to YouTube), and we have compared the results to those of a classifier trained using about 1600 manually labelled tweets.[1] The classifier trained via distant supervision turns out to yield substantially better classification accuracy than the one trained on manually annotated data.

We have analysed the effectiveness of our classification approach in different circumstances, so as to measure its robustness across different dimensions. We have first investigated the consequences of training our classifier with different sizes of automatically labelled data; here, we have found that training it with only 50,000 examples still outperforms the classifier trained with the manually labelled data. We have then run an additional experiment in which we have considered a smaller number of coarser classes (only 4, obtained by thematically grouping the original 14 classes); this experiment has shown that our classifier still outperforms the one trained with the manually labelled data. In an additional experiment we have compared the classification approaches on a test set of tweets dating from a time period much later than the one in which the training examples originated; the goal of this test was to investigate the effect of social media topical drift on classification effectiveness. In this latter experiment our approach still achieved high performance, while a large drop in performance instead affected the classifier trained with the manually labelled data. Finally, we also tested our approach on a set of non-English (namely, Arabic) tweets, so as to study the language independence of our distant supervision approach. Solid classification performance was noticed also on the Arabic test set.

The main contributions of our study are thus the following:

1. Proposing a novel tweet classification method based on distant supervision, which automatically harvests crowdsourced labelled data for the purpose of classifying microblogs under broad, general-purpose classes.
2. Proving that labels can be usefully transferred across different social media, thereby reducing the need of expensive manual labelling effort when tackling media-specific classification tasks.
3. Investigating the effectiveness of the above approach across several scenarios, including (1) varying numbers of automatically harvested training examples, (2) different class granularities, (3) different degrees of tweet recency, and (4) different languages.
4. Providing to the research community a set of 3,128 tweets manually labelled according to 14 general-purpose classes, to be used as benchmark data for future research.[2]

[1] Ours is thus a *single-label multi-class* classification task, since each tweet is assigned exactly one out of a set of 14 available classes.

[2] The dataset is available for download at http://alt.qcri.org/~wmagdy/resources.htm.

The paper is structured as follows. In Sect. 2 we discuss related work in distant supervision and microblog classification. Section 3 describes our method in detail, discussing all the steps we have taken to generate our automatically harvested training set. Section 4 describes our experimental setup, while Sect. 5 presents a number of experiments in which we compare the accuracy that can be obtained by training on our automatically harvested tweets with the one obtainable by training on manually annotated tweets. Section 6 presents further experiments in which we deviate, in different directions, from the basic setup of Sect. 4, to test the robustness of our approach to changing scenarios (namely: fewer training data, fewer and coarser classes, different languages). Section 7 concludes, sketching avenues for future research.

2 Related work

Distant supervision has been proposed in the literature for various applications, such as sentiment classification (Go et al. 2009; Marchetti-Bowick and Chambers 2012), relation extraction (Mintz et al. 2009), topical classification of blogs (Husby and Barbosa 2012), and tweet classification (Zubiaga and Ji 2013); this latter work will be discussed in detail later in this section. Most such works used distant supervision to obtain annotated data for their task from some other annotated dataset. For instance, Go et al. (2009) used the emoticons occurring in tweets as "silver" labels (i.e., as labels with more uncertain status that the ones found in usual "gold" standards) for tweet sentiment analysis. For relation extraction, Mintz et al. (2009) used textual features extracted from Freebase relations to train a relation classifier; Husby and Barbosa (2012) also used Freebase to obtain labels of Wikipedia articles, and used them for blog post classification by topic.

Previous work on microblog classification can be grouped according to three main dimensions: (1) the classification scheme used to classify the tweets, (2) the training method (e.g., standard supervision, distant supervision, etc.), and (3) the learning algorithm. As for (1), most published work for microblog classification focuses on classes targeted to a specific application. Genc et al. (2011) proposed a Wikipedia-based classification approach, by mapping tweets to the most similar Wikipedia pages; however, they tested their approach only on about 100 tweets grouped according to three events that occurred at the time of collecting the data. Kinsella et al. (2011) defined ten classes (e.g., Musicians, Photography, Soccer, MartialArts, Motors) for classifying blog posts. They used hyperlinks mentioned in the posts that link to webpages, and use the webpage metadata for classifying the post. The metadata includes page title, description, tags, and categories, whenever any of them are available. They showed a substantial improvement in classification accuracy when using metadata information. Classifying tweets is, however, a more challenging task than classifying blog posts, because of the tweets' limited short sentence length. Sriram et al. (2010) applied tweet-specific features in conjunction with bag-of-words to classify tweets into five broad classes (News, Events, Opinions, Deals, PrivateMessages). A simple classification task was discussed in Sankaranarayanan et al. (2009), where tweets were classified as News or Junk; a similar work appeared in Kothari et al. (2013), where tweets linking to news articles were classified as Comments or NewsReports. Also in Becker et al. (2011) the authors performed binary tweet classification, discriminating RealWorldEvents from NonRealWorldEvents. Irani et al. (2010) and Lee et al. (2011) studied tweet classification over trending topics. Lee et al. (2011) is the only work we are aware of that uses a fairly comprehensive set of classes (18), thereby covering a vast portion of the Twitter-sphere. However, these classes were motivated from trending topics on Twitter, which tend to change over time.

Most previous work on tweet classification by topic uses manually annotated training data (Becker et al. 2011; Chen et al. 2012; Irani et al. 2010; Kinsella et al. 2011; Kothari et al. 2013; Lee et al. 2011; Quercia et al. 2012; Sriram et al. 2010), which is both expensive and time-consuming. For training an effective classifier, a sizeable amount of training data is always required, especially when the number of classes is large. In addition, classifiers may need to be updated over time, so as to cope with concept drift, which may be especially severe in platforms as dynamic as those of social media. Therefore, methods that overcome the need for extensive manual annotation are to be preferred. Chen et al. (2012) apply a semi-supervised approach for classifying microblogs into six classes (which are a subset of the 14 classes used in our experimentation). They initially train a classifier with manually labelled data to probabilistically predict the classes for a large number of unlabelled tweets; then they train a new classifier also using the probabilistically predicted labels for the above-mentioned unlabelled tweets, and iterate the process to convergence. Zubiaga and Ji (2013) used distant supervision for tweet classification; as such, this work is highly relevant to the present work. Their approach consists in assuming that a tweet where a webpage URL occurs is on the same topic as that of the webpage; this is similar to our assumption about tweets mentioning YouTube links. The authors consider tweets linking to webpages classified under human-edited webpage directories. However, the shortcoming of their approach is that it depends on a human-edited directory which is limited in size and not necessarily up to date. Our proposed method is more

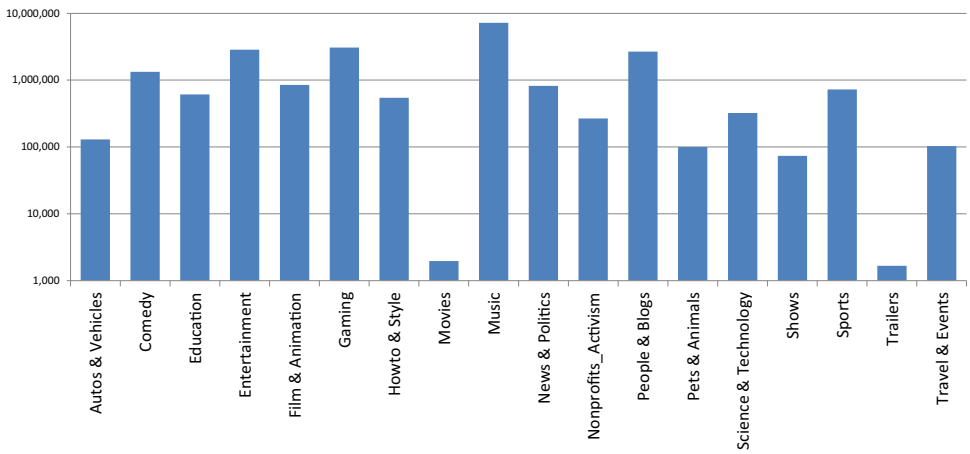

Fig. 1 Class distribution of the collected tweets

robust, since it is not dependent on any manually maintained resource.

Regarding learning algorithms, different ones have been used in the literature for the tweet classification task, the most common being Naïve Bayes (Kinsella et al. 2011; Sankaranarayanan et al. 2009; Sriram et al. 2010), decision trees (Irani et al. 2010; Lee et al. 2011), Labelled Latent Dirichlet Allocation (L-LDA) (Quercia et al. 2012), and support vector machines (SVMs) (Kothari et al. 2013; Lee et al. 2011).

Our work is different from the work reported in the literature in various respects. Considering the diversity of tweeted content, it is very hard to define classes for tweets that cover most of their aspects; we instead use standard classes from another social media platform. In addition, we propose a novel method for collecting automatically labelled data, to avoid the need for manually annotating training data. Our proposed method provides access to virtually unlimited amounts of free annotated data, amounts which can be increased at will essentially at no cost.

3 Leveraging automatically obtained labels for microblog classification

3.1 Harvesting labelled tweets

More than 4 million tweets in different languages linking to some YouTube video are tweeted everyday.[3] Every video on YouTube is assigned one of 18 predefined classes at the time of its upload. Our approach for collecting labelled tweets is based on the hypothesis that a tweet linking to a YouTube video can be reasonably assigned the same class that the video has been assigned. To validate this hypothesis, we have assigned labels to tweets linking to YouTube videos and used them to train a tweet classifier. We have used the Twitter API[4] with the string "youtube lang:en" to query the stream of English tweets with links to YouTube videos.[5] We have thus collected a set of ≈19.5 million tweets with hyperlinks to ≈6.5 million different YouTube videos in a period of 40 days between the end of March and the beginning of May 2014; it is often the case that multiple tweets link to the same video. We have then used the YouTube API[6] to extract the titles and classes of these videos, and have assigned these video classes as labels to the tweets linking them.

Figure 1 presents the distribution of the collected tweets across the 18 classes, plotted according to a log scale. As shown, the number of tweets per class ranges from only 1668 to more than 7 million. There are only three classes that contain less than 100k tweets (Movies, Trailers, and Shows). To avoid data sparseness, we have merged them with the class Film&Animation, since these three classes are topically similar. The class People&Blogs is the default class of YouTube, and is automatically assigned to videos when no class is specified by the user at the time of upload; we thus decided to drop this class, since we expect it to be noisy. Overall, these steps led to 14 classes with at least 100 k tweets per class.

We have noticed that the collected tweet set contains large number of retweets and duplicate tweets, which are tweets with the same text. We have thus filtered out all the tweets that are retweets or have duplicate text, so as to keep at most one occurrence of each tweet in the dataset; this has the effect of avoiding to train the classifier with repeated examples, which may lead to bias. Moreover, duplicate tweets often contain automatically generated text (e.g.,

[3] http://topsy.com/analytics?q1=site:youtube.com.

[4] http://twitter4j.org/en/index.html.

[5] This also captures tweets with shortened links to YouTube.

[6] http://developers.google.com/youtube/.

"Just watched video ..."), which can act as noise when training the classification model. This step reduced our dataset size from ≈19.5 million to ≈9.2 million tweets only. In the end, the smallest class in our data contains ≈62 k unique tweets.

3.2 Features, feature selection, and model generation

In the tweet classification literature various types of features have been used for training a classifier. These include Twitter-specific features (Kothari et al. 2013; Sriram et al. 2010), social network features (Lee et al. 2011), hyperlink-based features (Kinsella et al. 2011), and standard bag-of-words features, which are the most commonly used (Becker et al. 2011; Genc et al. 2011; Irani et al. 2010; Lee et al. 2011; Sankaranarayanan et al. 2009). Since feature design is not our main focus in this paper we simply apply a bag-of-words (BOW) approach, where each feature represents a term and the feature value is binary, denoting presence or absence of the term in the tweet. Nonetheless, in the following we discuss two methods for text enrichment that attempt to improve the performance of the BOW approach.

Since the length of tweets is very short and the information contained in them is thus limited, we have applied two different feature enrichment methods in an attempt to improve classification accuracy. The first method enriches the tweet text in the training data with the title of the linked video. This method is only applicable to our automatically obtained training tweets, since they all link to YouTube, but is not applicable in general to the unlabelled tweets we want to classify, since these may not link to any YouTube video. The second method duplicates the hashtags contained in the tweets and removes the hash character "#" from the second copy, so to allow the terms contained in the hashtags to increase the robustness of the term counts in the texts. In all our experiments, we applied simple text normalization, which includes case folding, elongation resolution (e.g., "cooooool" → "cool"), and hyperlinks filtration. Neither stemming nor stop word removal was applied.

We have then applied feature selection, by scoring all features via information gain (IG), defined as

$$\begin{aligned} \mathrm{IG}(t_k|c_i) &= H(c_i) - H(c_i|t_k) \\ &= \sum_{c\in\{c_i,\overline{c}_i\}} \sum_{t\in\{t_k,\overline{t}_k\}} P(t,c)\log_2\frac{P(t,c)}{P(t)P(c)} \end{aligned} \quad (1)$$

where $H(c_i)$ indicates the entropy of class c_i and $H(c_i|t_k)$ indicates conditional entropy; probabilities are evaluated on the space of training documents, where $P(t_k)$ and $P((\overline{t}_k))$ represent the fractions of tweets that contain the term t_k and do not contain t_k, respectively, and $P(c_i)$ and $P(\overline{c}_i)$ represent the fractions of tweets that are in class c_i and are not in class c_i, respectively. $\mathrm{IG}(t_k|c_i)$ measures the reduction in the entropy of c_i obtained as a result of observing t_k, i.e., measures the information that t_k provides on c_i.

All features are ranked according to their IG value for the class, after which a round-robin mechanism (Forman 2004) is applied in which the top n features are selected from each class-specific ranking and then merged to form the final feature space. In this way, for each class c_i the final set of selected features contains the n features that are best at discriminating c_i from the other classes, which means that all the classes in C are adequately championed in the final feature set.

We select the top 10,000 terms for each class; for 14 classes the theoretically maximum size of the feature space is thus 140,000 features, but the feature space is actually smaller since there is some overlap between the term sets selected for different classes.

4 Experimental setup

In our experimental setup we have focused on testing the effectiveness of our method at classifying generic tweets, regardless of the fact that they link or not to a YouTube video. We created two test sets,

- an automatically labelled test set, harvested in the same manner as our training set (the "silver standard"); and
- a manually labelled test set, consisting of tweets that do not necessarily have links to YouTube videos (the "gold standard").

4.1 Silver-standard training and test sets

From our dataset of automatically labelled tweets (described in Sect. 3.1) we randomly pick out for testing 1000 tweets for each class, for a total of 14,000 tweets evenly distributed across 14 classes. We refer to this test set as test_S (S standing for "silver"). We consider test_S as a "silver standard", since labels are not verified manually. For the rest of the automatically labelled tweets, we have opted to balance the number of tweets in each class by randomly selecting 100,000 tweets from each class, so as to match the number of tweets in the smallest class, namely Pets&Animals, which contains 98,855 tweets. The final training set thus contains ≈1.4 million tweets, which is three orders of magnitude larger than typical training sets used in the tweet classification literature. However, after applying duplicate and retweet filtering, as mentioned earlier, this number reduced to ≈913 k tweets (each class having 60 to 70 k

examples). We refer to this dataset as train_S. We trained SVMs on train_S using a linear kernel; this required a couple of hours on a standard desktop machine.

4.2 Gold-standard training and test sets

We created a second test set (the "gold standard") consisting of manually labelled generic tweets; this test set will henceforth be referred to as test_G (G standing for "gold"). There are two important reasons to have a manually labelled test set:

1. our test_S silver standard may be biased in favour of the system trained on train_S via distant supervision, because both datasets were sampled from the same distribution (i.e., they were labelled in the same automatic manner) and both consist of only tweets that link to YouTube; instead, the tweets in test_G do not necessarily contain a link to a YouTube video;
2. the manually labelled set test_G can be used for cross-validation experiments, in the manner described below. This will provide a solid baseline for the classifier trained using train_S.

To create a manually labelled set, it was difficult to randomly collect tweets covering all 14 classes, since some classes are rare and do not come up often in practice. To choose the tweets to label, we thus performed a guided search for each class using the Twitter API to stream tweets that contain hashtags similar to class names. This was done in the same month in which we collected our automatically labelled training dataset. For example, for the class Autos&Vehicles we collected tweets containing hashtags #autos or #vehicles. This helped us collect a set of tweets that, with high likelihood, had a substantial number of representatives for each of our classes of interest. We randomly selected 200 tweets for each class (based on hashtags), removed the hashtags that relate them with their possible class, and submitted them to a crowdsourcing platform for annotation. For every tweet, we asked at least three annotators if the displayed tweet matches the assumed class or not. Out of 2800 tweets representing 14 classes, only 1617 were assessed by all annotators as matching the assumed class; the number of tweets per class after validation ranged from 84 to 148. Examples of these tweets are shown in Table 3. This number of training examples is comparable to the numbers used in other studies from the literature (Becker et al. 2011; Genc et al. 2011; Irani et al. 2010; Kothari et al. 2013; Lee et al. 2011; Sankaranarayanan et al. 2009).

4.3 Classifiers

We have built the following classifiers for our experimentation:

- C_S: trained via distant supervision using train_S, which includes ≈913 k automatically labelled tweets.
- $C_{S(v)}$: same as C_S, with tweet enrichment using the title of the linked video.
- $C_{S(h)}$: same as C_S, with tweet enrichment obtained by adding the terms contained in the hashtags to the text.
- $C_{S(vh)}$: same as C_S, with tweet enrichment obtained by both heuristics above.

The S subscript indicates that all these classifiers have been trained on "silver" labels.

Further to this, we have run tenfold cross-validation (10FCV) experiments on the 1617 manually labelled tweets in test_G. We will then compare the results obtained by C_S and its variants on test_G, with the ones obtained by the classifiers generated in these 10FCV experiments; specifically, we will look at the results of

- C_G: this is not actually a single classifier but ten different classifiers, as generated within the 10FCV; that is, the results of applying C_G to test_G will be the union of the tenfolds, each of them classified within one of the ten experiments;
- $C_{G(h)}$: similar to C_G, but with tweet enrichment obtained by adding the terms contained in the hashtags to the text. Enrichment using the title of the linked video is not applicable, since most of the tweets in test_G do not link to YouTube.

Here, the G subscript indicates that all these classifiers have been trained on "gold" labels.

The main objective of our experiments was to examine if any of the C_S classifiers can achieve comparable (or even better) results with respect to the C_G classifiers, which would support our hypothesis and would also show the value of freely available labelled data. Different set-ups of the C_S classifier were examined for both test sets to find the optimal configuration that achieves the best results.

4.4 Evaluation

The evaluation measures we used in this task are "macroaveraged" precision (P), recall (R), F_1 (popularly known as the "F-measure"), and accuracy (A). That is, all of these measures were calculated for each class separately, after which the average was computed across the 14 classes. Since our test sets contain fairly balanced numbers of examples from each class

- these macroaveraged figures are very similar to the corresponding "microaveraged" ones (where classes more frequent in the test set weigh more), which are then not reported explicitly;

– accuracy is indeed a reasonable measure of classification effectiveness (this is unlike the cases of severe imbalance, when accuracy is unsuitable).

5 Experiments

5.1 Choosing the learning algorithm

As a first step we performed preliminary experiments aimed at finding a good learning algorithm for our task. We run these preliminary experiments using a small training set consisting of 2000 training examples per class randomly selected from train_S, and using test_S for testing.

We tested three different learning algorithms: (a) a distance-weighted k-NN learner (Yang and Liu 1999; b) a multinomial Naïve Bayesian learner (McCallum and Nigam 1998), also from the WEKA suite; (c) a linear-kernel support vector machine (SVM) learner (Joachims 2002), in Thorsten Joachims' SVM-light implementation.[7]

As we have noted in the introduction, ours is a single-label multi-class task (i.e., exactly one class out of the 14 available classes must be attached to each tweet). The k-NN and Naïve Bayesian methods are "natively" single-label multi-class, which suits our task well. Instead, SVMs are binary in nature; we thus used the "multiclass" option[8] available in SVM-light, which optimally converts the results of 14 independent binary classifiers into a single-label decision.

For k-NN, we tested all values of $k \in \{1, 2, \ldots, 9, 10, 20, \ldots, 90, 100\}$; the best result was $A = 0.425$, obtained for $k = 6$. For the Naïve Bayesian learner we instead obtained a very low $A = 0.255$. For the SVM learner, we tested all values of $c \in \{10^0, 10^1, \ldots, 10^5\}$, where c is the parameter that sets the tradeoff between model complexity and training error. The best result was $A = 0.543$, obtained for $c = 10^5$.

Since in these preliminary experiments the SVM learner was by far the best performer, we will use it as our learning algorithm in the rest of the paper, with the c parameter set to 10^5. An additional benefit of using the SVM learner is that, in the above preliminary experiments, it proved by far the fastest.

5.2 Results

Tables 1 and 2 report the classification results obtained on the "silver" test set test_S and on the "golden" test set test_G, respectively. All results in both tables display a relatively good effectiveness for a single-label 14-class classification task, where random classification would achieve an expected classification accuracy of $100/14 \approx 7.142\,\%$.[9]

Table 1 Classification results on the silver-standard test set (test_S)

	P	R	F_1	A
C_S	**0.583**	0.573	0.564	0.574
$C_{S(v)}$	0.574	0.567	0.560	0.568
$C_{S(h)}$	0.582	**0.575**	**0.568**	**0.576**
$C_{S(vh)}$	0.576	0.569	0.562	0.571

Boldface indicates the best performer

Table 1 shows that the "enhanced" setups of the C_S classifier did not lead to noticeable improvement. Enriching the training tweets with the title of the linked video even led to a small degradation in performance, while enriching the representation of the tweets by duplicating hashtags achieved only slightly better results.

The results in Table 1 suggest that our idea of using YouTube labels for training a tweet classifier is a reasonable one. Nevertheless, the main experiments are those reported in Table 2, which reports results obtained on a truly gold standard.

Table 2 reports the results of different setups of C_S and C_G on test_G. All different setups of C_S achieved better performance than all different setups of C_G, which confirms that our method for inexpensively acquiring large numbers of automatically annotated training examples is more effective than the (more expensive) method of labelling a limited number of training examples.

Regarding the best setup for the training data, we noticed that hashtag term duplication improved the performance of C_G over all scores, but did not lead to any improvement in the case of C_S. The limited number of training examples used for generating C_G can be the reason for this result: here some enrichment to the representation of the training examples seems to help, unlike in the case of C_S, which was trained via a large number of training examples and does thus not require further enrichment. The best result achieved for C_S and its variants was $A = 0.611$ and $F_1 = 0.579$ (which was obtained for C_S itself), which is substantially higher than the best result achieved for C_G

[7] http://svmlight.joachims.org/.

[8] https://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html.

[9] These results are, relative to the difficulty of the task, only deceptively inferior to other results published in the tweet classification literature. For instance, in Lee et al. (2011) (which, as mentioned in Sect. 2, is the only article in the tweet classification literature that deals with a set of classes comparable to ours), the authors obtain 70.96 % accuracy. However, their dataset is easier than ours: in their case, 70.96 % accuracy is 3.68 times higher than their trivial acceptor (the classifier that always picks the majority class), while our 0.574 accuracy value is 8.03 times higher than that obtained by our trivial acceptor.

Table 2 Classification results on the gold-standard test set ($test_G$)

	P	R	F_1	A
C_G	0.511	0.506	0.507	0.518
$C_{G(h)}$	0.541	0.534	0.537	0.546
C_S	**0.619**	**0.588**	**0.579**	**0.611**
$C_{S(v)}$	0.570	0.566	0.548	0.586
$C_{S(h)}$	0.600	0.583	0.573	0.605
$C_{S(vh)}$	0.578	0.567	0.551	0.588

Boldface indicates the best performer

and its variants ($A = 0.546$ and $F_1 = 0.537$, which was obtained for $C_{G(h)}$). From now on, C_S and $C_{G(h)}$ will be used when comparing the distant supervision and the standard supervision approaches in further experiments. Anyway, the above result validates our hypothesis that classification labels from YouTube video could be applied to tweets linking them, and used to train a tweet classifier that is more effective than one obtained by manually labelling training data.

Table 3 shows the complete confusion matrix obtained using the classifier C_S on $test_G$. The classifier achieved a F_1 value higher than 0.650 in classifying Autos&Vehicles, Gaming, HowTo&Style, Pets&Animals, Sports and Travel&Events. We further analysed the results of classes with a poor F_1 value. It is clear from the table that most of such classes were confused with the class News&Politics. Other confusion between classes occurred, quite obviously, between related classes such as Film&Animation and Entertainment, and Comedy and Entertainment. In some cases, this may be due to the multifaceted nature of a tweet that may naturally refer to more than one class. Examples of this phenomenon are the wrongly classified examples presented in Table 4, where e.g., the tweet "Thief Attacks Victim on Scooter" is classified as News&Politics instead of as Autos&Vehicles. Both classes might be correct based on the content of the tweet. The examples presented in Table 3 show that some of the tweets can actually be classified into more than one class. This can motivate exploring multi-label multi-class classification in future work.

6 Further experiments

In this section

1. we further investigate the robustness of our approach by measuring the consequences on classification effectiveness of increasing/decreasing the amount of training data;
2. we examine the performance of classification using distant supervision when using a smaller number of coarser classes;
3. we test how robust the classifiers trained on automatically labelled data are with respect to concept drift;
4. we examine how language independent our approach is by performing a classification experiment on non-English tweets (Arabic, in our case).

6.1 Effect of training data size on classification accuracy

In the previous section we have shown that, when compared on the same test set $test_G$, C_S (the best of the classifiers trained via distant supervision, i.e., on silver labels) achieved substantially better results than $C_{G(h)}$ (the best of the classifiers trained on gold labels); this happened when using ≈913 k training examples with C_S vs. only 1617 for $C_{G(h)}$. Even though coming up with a dataset of ≈913 k automatically labelled examples is much cheaper than coming up with one of 1617 manually annotated ones, it is interesting to study the effect of reducing the number of automatically annotated examples so as to see to what extent the automatically labelled data would retain its advantage. In addition, we have also examined the consequences of using more automatically annotated training examples, so as to see if there are further margins of improvement.

Figure 2 shows a log-scale plot of classification accuracy as a function of the amount of silver training data. The dotted horizontal line represents the accuracy achieved by $C_{G(h)}$ using the 1617 manually labelled training examples. As shown, C_S continues to outperform $C_{G(h)}$ when as few as ≈50 k training examples are used; note that 50 k tweets linking to YouTube videos covering all the classes could be easily collected in 1 day. However, with fewer than 50 k automatically labelled training examples the performance of $C_{G(h)}$ is higher than that of C_S. When using the same small number of training examples (1617), the accuracy of C_S is less than half the accuracy of $C_{G(h)}$. This highlights the fact that, as expected, YouTube-derived labels are not of the same quality as manually obtained ones. It thus makes sense, when using automatically derived labels, to use large numbers of them, especially since they come at essentially no cost.

We further tested the effects on classification accuracy of increasing the size of the training set even beyond 1.4 m (which is the size of $train_S$ above); note however that this has the effect of disrupting the almost perfect balance among the classes, since (as previously mentioned) some classes had no more than 100k examples in our crawl. As shown in Fig. 2, when increasing the size of the training data beyond 1.4 m, accuracy slightly increased inasmuch as the imbalance was limited to the largest class having double the examples of the smallest class. However, when

Table 3 The confusion matrix for the classifier $C_{S(h)}$ as tested on test_G

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Autos&Vehicles	136	0	0	1	0	0	0	1	4	0	1	2	1	2
2	Comedy	3	33	1	7	8	3	3	8	2	2	11	4	9	7
3	Education	0	0	23	0	1	2	4	2	20	3	0	33	5	1
4	Entertainment	2	3	1	38	3	5	9	18	10	3	2	3	17	3
5	Film&Animation	4	1	4	5	55	2	3	5	8	2	3	6	5	13
6	Gaming	1	1	0	2	3	105	3	3	3	0	4	12	6	1
7	HowTo&Style	4	1	0	3	4	5	86	2	0	1	9	4	5	1
8	Music	1	0	1	1	5	0	4	55	2	2	1	2	5	5
9	News&Politics	8	1	2	2	1	1	2	7	56	1	2	10	4	5
10	Nonprofits&Activism	4	0	3	2	0	3	4	1	18	30	4	8	6	5
11	Pets&Animals	0	0	0	1	0	2	0	0	2	0	105	2	3	1
12	Science&Technology	2	0	3	1	2	4	3	2	16	3	1	69	3	4
13	Sports	8	1	0	4	0	2	1	3	5	0	2	0	99	5
14	Travel&Events	4	0	1	0	2	1	8	5	6	1	4	6	3	98
	Precision	0.77	0.80	0.59	0.57	0.65	0.78	0.66	0.49	0.37	0.63	0.70	0.43	0.58	0.65
	Recall	0.92	0.33	0.24	0.32	0.47	0.73	0.69	0.65	0.55	0.34	0.91	0.61	0.76	0.71
	F_1	0.84	0.46	0.35	0.41	0.55	0.75	0.67	0.56	0.44	0.44	0.79	0.50	0.66	0.68

Table 4 A few examples of tweets misclassified by $C_{S(h)}$

Tweet	True label	Predicted label
Female Softball Player Comes Out #CelebrityNews #Funny #Funny News #Jokes http://t.co/K92JGuDArC	Comedy	Sports
Thief Attacks Victim on Scooter	Autos&Vehicles	News&Politics
RT @Britt Coletti: State adopts new teacher	Education	News&Politics
I learn #German on my iPhone - just amazingly cool and only 99 cent http://t.co/AwrsvfkLb8 #ios #cool	Education	Science&Technology

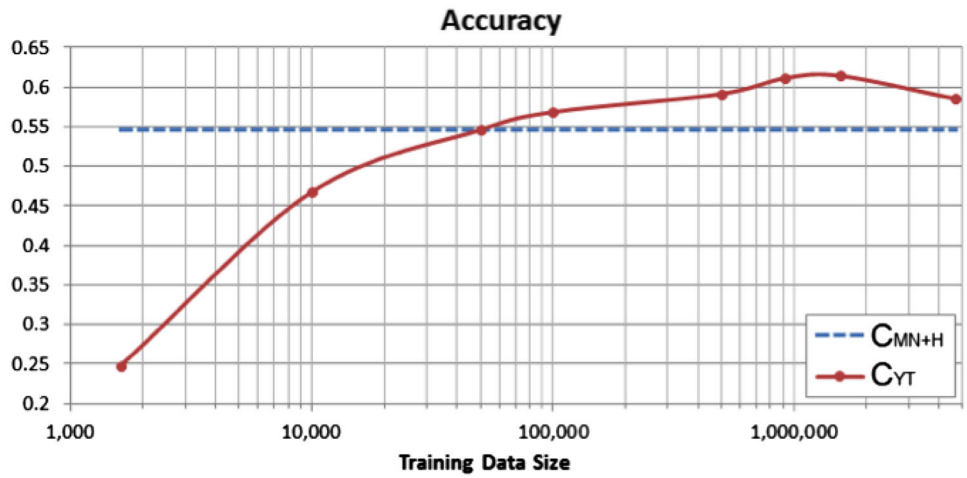

Fig. 2 Classification accuracy as a function of the amount of training data

the level of imbalance went beyond that, accuracy suffered despite the larger size of the training set.

6.2 Testing distance supervision with smaller numbers of classes

Our experiments so far had to do with classifying tweets into 14 mid-grained classes. A set of coarse classes can easily be extracted from the collected data. This increases the usability of the data for applications that require general classes. To perform this, we have thematically grouped our 14 classes into only 4 classes. We have grouped Education with Science&Technology, News&Politics with Nonprofits&Activism, Autos&Vehicles with Sports, and the remaining classes Comedy, Film&Animation, Gaming, HowTo&Style, Music, Pets&Animals, Travel&Events with Entertainment. We have then retrained both $C_{G(h)}$ and C_S using the new classification scheme. The results obtained on test_G are shown in Table 5. As shown, C_S continues to achieve superior performance with respect to $C_{G(h)}$, which further illustrates the effectiveness of distant supervision.

6.3 Effects of concept drift on classification effectiveness

One of the main characteristics of social media, and of Twitter in particular, is its highly dynamic nature, since the topics discussed change dramatically over time (Magdy and Elsayed 2014); as a consequence, the characteristics of tweets that belong to a certain class also tend to change, a phenomenon that in machine learning is called *concept drift* (Sammut and Harries 2011). As a consequence, a model trained for a given tweet classification task could become less effective over time. To ascertain to what extent this problem affects our distant supervision method, we have carried out experiments to ascertain how much effectiveness drops when models trained by distant supervision are tested on tweets harvested several months after the models were trained. Most literature on tweet classification has so far neglected studying the consequences of concept drift.

In December 2014 (i.e., 8 months after collecting all the data discussed in the previous sections) we have thus collected another set of tweets. Hashtags of class names were used to collect the tweets, then a random set of 200 tweets was selected from each class and annotated by crowdsourcers according to the same method used for creating test_G. Out of the 2800 tweets, only 1511 were assessed by the annotators to be matching the assumed class. We call the resulting test set test_{G_2}. We applied our two classifiers $C_{G(h)}$ and C_S to the new test set test_{G_2}; results are reported in Table 6.

As shown in Table 6, $C_{G(h)}$ suffered from a significant drop in performance, while C_S obtained on test_{G_2} results comparable to those obtained on test_G. This seems to suggest that one of the disadvantages of using a small number of manually labelled examples to train a tweet classification model is a drop in effectiveness over time due to the drift in social media content, which requires a robust model trained on a wide range of examples. Our findings point to another advantage of our distant supervision approach for tweet classification.

6.4 Experiments on non-English content

One of the main advantages of our approach is that it is language independent, since no language-specific processing is required. Our final experiment thus concerned the application of our distant supervision approach to a language for which much fewer classification studies are available, i.e., Arabic. We collected a set of Arabic tweets linking to YouTube by running the query "youtube lang:ar" on the Twitter API. We collected more than 5 million tweets; the minimum number of tweets per class was 35,460 (for class Pets&Animals). We extracted 1000 tweets at random from each class for creating a "silver" test set, and selected from the remaining ones a balanced set of tweets to be used as "silver" training data. The final size of the training set was ≈482 k, representing 14 classes. We attempted to use the same methodology of using hashtags for creating a manually labelled test set, but unfortunately the class names, once translated into Arabic, did not match enough tweets. Therefore, in this analysis we only rely on the "silver" test set only, which was shown in our earlier experimentation to be a good indicator of classification performance.

Table 5 Classification results using 4 coarser classes instead of the 14 original ones

	P	R	F_1	A
$C_{G(h)}$	0.593	0.588	0.590	0.699
C_S	**0.710**	**0.701**	**0.705**	**0.787**

Boldface indicates the best performer

Table 6 Classification results on a test set of tweets collected 6 months later after the tweets used for training

	P	R	F_1	A
$C_{G(h)}$	0.462	0.456	0.450	0.465
C_S	**0.615**	**0.595**	**0.587**	**0.611**

Boldface indicates the best performer

Table 7 Classification results on the Arabic silver-standard test set

	P	R	F_1	A
C_S	0.644	0.646	0.640	0.646

We applied one of the available tools for social Arabic text normalization (Darwish et al. 2012), which performs character normalization, word elongation resolution, and emotion detection. Normalized Arabic tweets were then used to train our classifier; as before, a BOW approach using IG for feature selection was used.

Table 7 shows the results of classifying the Arabic tweet dataset. The results obtained are even higher than those on the English data, which illustrates the effectiveness of our distant supervision method regardless of the language in which tweets are expressed.[10]

One interesting finding is that, as we noticed when extracting the meta-information from the videos linked by the Arabic tweets, 7 % of these videos have their title and description in a Latin-script language (mostly English). This shows that this approach could be applied even to languages with low resources on YouTube, since tweets in one language can link to videos titled in a different, resource-richer language.

7 Conclusion

In this paper, we have experimentally demonstrated the effectiveness of a "distant supervision" approach to tweet classification, consisting in automatically obtaining labelled data from one social media platform (YouTube) and using this data for training a classifier for another such platform (Twitter). Our proposed distant supervision method generates a large amount of freely available labelled training data, thus overcoming the need for manual annotations. As a side result, we have also generated a dataset of 3128 annotated tweets (the union of test_G and test_{G_2}) that we make available to the research community.

When comparing the quality of a classifier trained via our distant supervision method with the one of a classifier trained on ≈1.6 k manually labelled tweets, we have shown that the former outperforms the latter when only ≈50 k examples are used for training, which can be easily collected in one day using the freely available Twitter API. Our classification technique also showed superior effectiveness over the traditional one even when a smaller number of more general classes were considered instead. In addition we showed the robustness of our approach once used on resource-poor languages, and its robustness with respect to time drift.

For future work, it would be interesting to apply advanced pruning and data cleaning approaches for our collected training data, since it is collected automatically and is thus prone to noise; data cleaning could potentially improve performance even further. In addition, it would be interesting to apply *transfer learning* (TL) (Pan and Yang 2010; Pan et al. 2012; Raina et al. 2007) to use our labelled tweets for different classification schemes. TL focuses on alleviating the need of labelling examples for a "target domain" by leveraging training examples from a different (although related) "source domain" for which the amount of available labelled examples is higher. TL allows making use of these examples to train an effective classifier for the target domain, thus allowing to diminish or completely remove the cost involved in the manual generation of training documents (Do and Ng 2005). In our case, this might be useful for sentiment analysis and emotion classification, since tweets of classes Entertainment and Comedy are more likely to be good indicators of positive emotions, while classes such as News&Politics sadly tend to have the opposite polarity.

References

Becker H, Naaman M, Gravano L (2011) Beyond trending topics: real-world event identification on Twitter. In: Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM 2011). Barcelona, ES

Bollen J, Mao H, Zeng XJ (2011) Twitter mood predicts the stock market. J Comput Sci 2(1):1–8

Chen M, Jin X, Shen D (2011) Short text classification improved by learning multi-granularity topics. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011). Barcelona, ES, pp 1776–1781

Chen Y, Li Z, Nie L, Hu X, Wang X, Chua TS, Zhang X (2014) A semi-supervised Bayesian network model for microblog topic classification. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012). Mumbai, IN, pp 561–576

Darwish K, Magdy W, Mourad A (2012) Language processing for Arabic microblog retrieval. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012). Maui, US, pp 2427–2430

[10] The fact that we obtain better results on Arabic than on English might at first seem surprising, but is not implausible. The literature on multilingual classification (see e.g., Gonçalves and Quaresma 2010) reports many cases in which substantially different levels of accuracy are obtained for different languages, even when the training data and the test data are exactly the same (i.e., each training/test document is a translation equivalent of a training/test document in another language). These differences may be due to a multiplicity of factors, including the different accuracy of preprocessing tools (e.g., stop word lists, stemmers, lemmatizers, decompounders, parsers, etc.) in the different languages, the presence of different linguistic phenomena in different languages, etc. In our case, the documents are not even translation equivalents of each other, so the difference should be even less surprising.

De Choudhury M, Diakopoulos N, Naaman M (2012) Unfolding the event landscape on Twitter: Classification and exploration of user categories. In: Proceedings of the 15th ACM Conference on Computer Supported Cooperative Work (CSCW 2012). Seattle, US, pp 241–244

Do CB, Ng AY (2005) Transfer learning for text classification. In: Proceedings of the 19th Annual Conference on Neural Information Processing Systems (NIPS 2005). Vancouver, CA, pp 299–306

Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. PLoS One 6(12)

Forman G (2004) A pitfall and solution in multi-class feature selection for text classification. In: Proceedings of the 21st International Conference on Machine Learning (ICML 2004). Banff, CA, pp 38–45

Genc Y, Sakamoto Y, Nickerson JV (2011) Discovering context: Classifying tweets through a semantic transform based on Wikipedia. In: Proceedings of the 6th International Conference on Foundations of Augmented Cognition (FAC 2011). Orlando, US, pp 484–492

Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. Stanford University, Tech. rep

Gonçalves T, Quaresma P (2010) Polylingual text classification in the legal domain. Informatica e Diritto XIX(1–2), pp 203–216

Husby SD, Barbosa D (2012) Topic classification of blog posts using distant supervision. In: Proceedings of the EACL Workshop on Semantic Analysis in Social Media. Avignon, FR, pp 28–36

Imran M, Castillo C, Diaz F, Vieweg S (2014) Processing social media messages in mass emergency: a survey. http://arxiv.org/abs/1407.7071v2

Irani D, Webb S, Pu C, Li K (2010) Study of trend-stuffing on Twitter through text classification. In: Proceedings of th 7th Conference on Collaboration, Electronic Messaging, Anti-Abuse and Spam (CEAS 2010). Redmond, US

Joachims T (2002) Learning to classify text using support vector machines: methods, theory and algorithms. Kluwer Academic Publishers, Dordrecht

Kinsella S, Passant A, Breslin JG (2011) Topic classification in social media using metadata from hyperlinked objects. In: Proceedings of the 33rd European Conference on Information Retrieval (ECIR 2011). Dublin, IE, pp 201–206

Kothari A, Magdy W, Darwish K, Mourad A, Taei A (2013) Detecting comments on news articles in microblogs. In: Proceedings of the 7th International Conference on Weblogs and Social Media (ICWSM 2013). Cambridge, US

Lee K, Palsetia D, Narayanan R, Patwary MMA, Agrawal A, Choudhary A (2011) Twitter trending topic classification. In: Proceedings of the 6th Workshop on optimization-based techniques for emerging data mining problems (OEDM 2011). Vancouver, CA, pp 251–258

Magdy W, Elsayed T (2014) Adaptive method for following dynamic topics on Twitter. In: Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM 2014). Ann Arbor, US

Marchetti-Bowick M, Chambers N (2012) Learning for microblogs with distant supervision: Political forecasting with Twitter. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012). Avignon, FR, pp 603–612

McCallum AK, Nigam K (1998) A comparison of event models for naive Bayes text classification. In: Proceedings of the AAAI Workshop on Learning for Text Categorization. Madison, US, pp 41–48

Mintz M, Bills S, Snow R, Jurafsky D (2009) Distant supervision for relation extraction without labeled data. In: Proceedings of the 47th Annual Meeting of the ACL and 4th International Joint Conference on Natural Language Processing (ACL/IJCNLP 2009). Singapore, SN, pp 1003–1011

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359

Pan W, Zhong E, Yang Q (2012) Transfer learning for text mining. In: Aggarwal CC, Zhai C (eds) Mining text data. Springer, Heidelberg, DE, pp 223–258

Quercia D, Askham H, Crowcroft J (2012) TweetLDA: Supervised topic classification and link prediction in Twitter. In: Proceedings of the 4th ACM Conference on Web Science (WS 2012). Evanston, US, pp 247–250

Raina R, Battle A, Lee H, Packer B, Ng AY (2007) Self-taught learning: transfer learning from unlabeled data. In: Proceedings of the 24th International Conference on Machine Learning (ICML 2007). Corvalis, US , pp 759–766

Sammut C, Harries M (2011) Concept drift. In: Sammut C, Webb GI (eds) Encyclopedia of Machine Learning. Springer, Heidelberg, pp 202–205

Sankaranarayanan J, Samet H, Teitler BE, Lieberman MD, Sperling J (2009) TwitterStand: news in tweets. In: Proceedings of the 17th ACM International Conference on Advances in Geographic Information Systems (GIS 2009). Seattle, US, pp 42–51

Sriram B, Fuhry D, Demir E, Ferhatosmanoglu H, Demirbas M (2010) Short text classification in Twitter to improve information filtering. In: Proceedings of the 33rd ACM International Conference on Research and Development in Information Retrieval (SIGIR 2010). Geneva, CH, pp 841–842

Yang Y, Liu X (1999) A re-examination of text categorization methods. In: Proceedings of the 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR 1999). Berkeley, US, pp 42–49

Zubiaga A, Ji H (2013) Harnessing Web page directories for large-scale classification of tweets. In: Posters Proceedings of the 22nd International World Wide Web Conference (WWW 2013). Rio de Janeiro, BR, pp 225–226