2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

ClassStrength: A Multilingual Tool for Tweets Classification

Walid Magdy* and Mohamed Eldesouky†
*School of Informatics, The University of Edinburgh, UK
†Qatar Computing Research Institute, HBKU, Doha, Qatar
wmagdy@inf.ed.ac.uk, mohamohamed@hbku.edu.qa

***Abstract*—In this paper we present our multilingual tweet classification tool. *ClassStrength* provides a set of classification models in different languages that classify tweets into 14 general-purpose categories, including: sports, politics, entertainment, comedy, etc. Our classifier uses a distant-supervision approach for creating training data in any available language on Twitter. The classifier uses a soft-classification scheme, where it generates a likelihood score for a tweet to match each of the 14 categories. The initial version of our tool covers five languages, namely: English, Arabic, French, German, and Russian. More languages are to be covered in next releases. The classification model created for each language is generated from hundreds of thousands of training tweets. Our evaluation to the classifier shows superior accuracy compared to standard manual methods. Our reported accuracy is 84% based on crowd preferences over a balanced test set of English tweets covering all 14 classes.**

I. Introduction

Interest in classifying social posts has increased with the wide spread of social websites, such as Facebook and Twitter. Classifying short social posts such as tweets has been investigated for different applications, including sentiment analysis [12], news detection [5], [10], and general-purpose categorization [4], [11]. Classifying social posts into general-purpose categories has various applications, such as categorized search, social media analysis, user profiling, and recommendation systems. In addition, general categories could still be used for other classification schemes through transfer learning [9].

Most of previous work on tweets classification focused on preparing a set of manually labeled tweets to train a classifier for the targeted classification task [1], [3], [5], [8], [10]. Data annotation is an expensive task in terms of both time and money, especially when a large number of classes is used, since a sufficient number of examples are required to effectively model each class.

One of the main characteristics of social media is its high connectivity, even across different social networks. In this work we utilize the network links among different social platforms for automatically driving labels for the purpose of building classification models. The process is called distance-supervision [7], [6]. This method does not require the preparation of any manual annotation. It uses crowd-sourced labels from one social domain, namely YouTube, to train a classifier for tweet classification. The advantage of this method stems from the ubiquitous availability of free training instances for automatic classification. Also, standardized categories are adopted and used in YouTube.

The approach collects a large set of tweets, in a given language, linking to YouTube videos. Each YouTube video is assigned to one category out of 18 predefined general-purpose categories. The category is assigned by users when uploading a video to YouTube. The approach transfers video categories as labels to the tweets linking them. This creates a large set of automatically labeled tweets regardless of the language. The collected set of labeled tweets is then used to train a classifier after a set of preprocessing and pruning steps to reduce noise and normalize text. Previous work [7], [6] has shown that the proposed approach for collecting large number of automatically-labeled tweets to train a classifier achieves significantly better results compared to using a small set of manually labeled tweets. Analysis in [6] shows that a set of 50,000 automatically-labeled tweets used to train a classifier would outperform a set of 1,600 carefully manually labeled tweets. In spite of this large difference in the size of training data of both methods, automatically collecting 50,000 tweets linking to YouTube is significantly more efficient than manually labeling a set of 1000 tweets in both cost and time.

In this paper, we offer our tweets classification tool *ClassStrength*[1], which applies distant-supervision algorithm to collect labels automatically for training tweet classifiers in 5 different languages. For each of the languages, hundreds of thousands of training data were collected and used for training a classification model. These models are then used by our classifier. *ClassStrength* is composed of mainly three components:

1) Feature extraction tool, which is responsible of extracting features from tweets that would be used by our classifier. This component contains the feature extraction script, in addition to, the feature vector files, where there is a vector file for each language.

[1] **http://alt.qcri.org/class_strength/**

ASONAM '17, July 31 -August 03, 2017, Sydney, Australia

ACM ISBN 978-1-4503-4993-2/17/07?/$15.00
http://dx.doi.org/10.1145/3110025.3110162

2) Classification models, are the models used by the classifier to calculate the likelihood probabilities of classifying a tweet into each of the possible general-purpose classes. There is a model file for each language.
3) Language detection module, which is responsible for detecting the language of a tweet before forwarding it to the correct classification model. This component is optional, since user can specify the language manually.

In the rest of the paper, we describe the components of our tool in details, including the collection process of the training data and the classifier performance.

II. Training Data

We collected a set containing millions of tweets in 5 different languages that are linked to YouTube videos. We used the Twitter API[2] with the string "youtube lang:xx" to query the stream of tweets in a given language ("xx") with links to YouTube videos[3]. We collected in one month a set of 6M to 20M tweets in each language, that have a link to YouTube videos covering 18 categories.

The large amount of data collected contained some noise and duplicates. In order to clean up the data for the classifiers' training phase, we applied the following pre-processed steps:

- **Shallow text normalization**, simple cleaning and normalization is applied to the tweet's text, including case-folding (for Arabic, we use normalization described in [2]), Hyperlinks removal (mostly are YouTube links), tokenization, and punctuation removal.
- **Removing duplicates**, retweets and tweets with exact content are filtered out. This helps to avoid over-fitting for tweets that receive large number of retweets. In addition, with this step, automated tweets with fixed text phrases would be mostly filtered out, e.g. "*I liked a video on YouTube: ...*".
- **Filtering and Merging Categories**; One of the largest categories in YouTube is People & Blogs. However, this category is the most noisy category on YouTube, since it is the default category that gets assigned to a video in case the user did not assign any. Therefore, we decided to remove tweets carrying this label from our set. Furthermore, we noticed that categories Film & Animation, Movies, Trailers, and Shows are the least popular ones in all the languages. Thus we merged them all into one category since they are all related. This end up our training data to contain **14** categories only.
- **Balancing Data**, as shown in [6], collected labels from YouTube are highly unbalanced. For example, Music is usually the most popular category linked on Twitter. In this step, training data are balanced by reducing the size of training samples from each category to the size of the smallest one. This insures training a less-biased classifier.

The previous steps reduces the amount of training samples significantly. Nonetheless, the amount of data we end up with is still large, and more stream of tweets could still be used to add more training data. Table I shows the size of training data set used for training the classification model of each language.

Language	Size of balanced training data
English	913K
Arabic	482K
French	544K
German	349K
Russian	405K

TABLE I: Number of tweets used to train the classification model of each language

III. Classification Algorithm

A. Classifier

Due to the humongous size of our training data, we simply applied 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.

SVM mulitclass classifier [4] is used to train our classification model for each language separately using a linear kernel. SVM was shown to be the most effective and efficient among other classification methods in [6].

B. Performance

An objective evaluation was conducted in [6] showing that this classification approach using distant-supervision is superior over standard approach that uses limited number of manually labeled data, where the first approach was reported to achieve an accuracy of 61% compared to only 53% to the latter.

The evaluation in [6] assumes a hard classification task, where each tweet can only be assigned to only one label. In fact, this is not the case in many situations. Thus, *ClassStrength* is designed to assign a weight for each class according to its suitability to the tweet's text. To measure the performance of our tool, we applied an additional evaluation to the classifier. Since it assumes a soft-classification task, where multiple categories can apply to the same tweet, we used the same test set in [6] that contains 1677 English tweets and applied *ClassStrength* to it. We then presented the top predicted class for each tweet to workers on Crowdflower[5] and asked how they evaluate the classification of the tweets based on the predicted category. The workers were allowed to choose among three different choices: 1) Perfect: the category matches the tweet's content perfectly; 2) Acceptable: the predicted category is OK, but not perfect; or 3) Bad: the predicted category has no relation to the tweet. We asked three annotators to label each tweet, and majority voting was used. The agreement among annotators was 65%.

Figure 1 shows the distribution of answers of annotators. As shown, 56% of the tweets were seen to be perfectly classified,

[2]http://twitter4j.org/en/index.html
[3]This also captures tweets with shortened links to YouTube.
[4]https://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html
[5]https://crowdflower.com

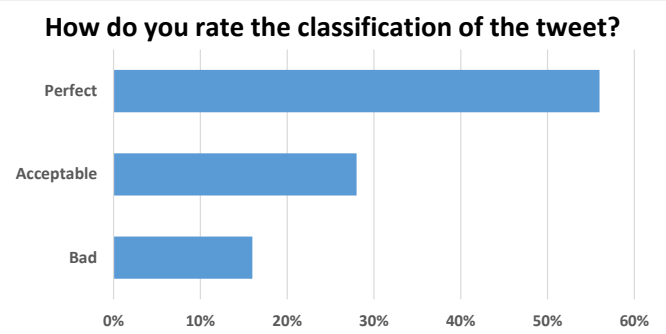

Fig. 1: Classes list of the two classification schemas existing in *ClassStrength*

while 28% were seen not perfect, but still acceptable. Only 16% of the tweets were seen to carry inaccurate category. This result shows that 84% of the time *ClassStrength* leads to a reasonable prediction for tweets category out of 14 different categories. This results are expected to generalize to other languages as well.

IV. Classification Tool

ClassStrength tool is available for free use and download at the following link: http://alt.qcri.org/class_strength/. The tool has three operational modes:

Online Mode: This is the mode used mainly for Demo purposes. On the tool webpage, a text box is provided for users to type in some text in any of the five languages, and a graph of the strength of each category of the 14 categories is displayed showing the prediction of classification of the typed text. The user can manually select the language of the text, or keep the default selection which automatically detects the language and apply the classification to it.

Figure 2 shows an example tweet in the text box and the displayed strength of each of the categories as estimated by *ClassStrength*. As shown, the tweet is about President Obama meeting an NBA basketball team at the White House. The language detector identified the language of the tweet, and then passed the text to the English classification model. As shown, the classifier detected that the main category of the tweet is News & Politics, while the Sports category also received some positive score. All the remaining categories received negative scores, showing that they do not apply to the tweet.

This mode allows users to type freely in the text box and get an illustration to the strength of each category of the input text. The best matching category to the input text receives a normalized score of 100%, while the remaining ones receive fractions of this, or negative scores.

Batch Mode: This mode allows users to submit a text file containing a list of tweets to be classified on *ClassStrength* server, and the output would be ready for download directly. If the user knows the language of the submitted tweets, it is better to select the language manually, and then the classification model of this language would be applied directly to the list

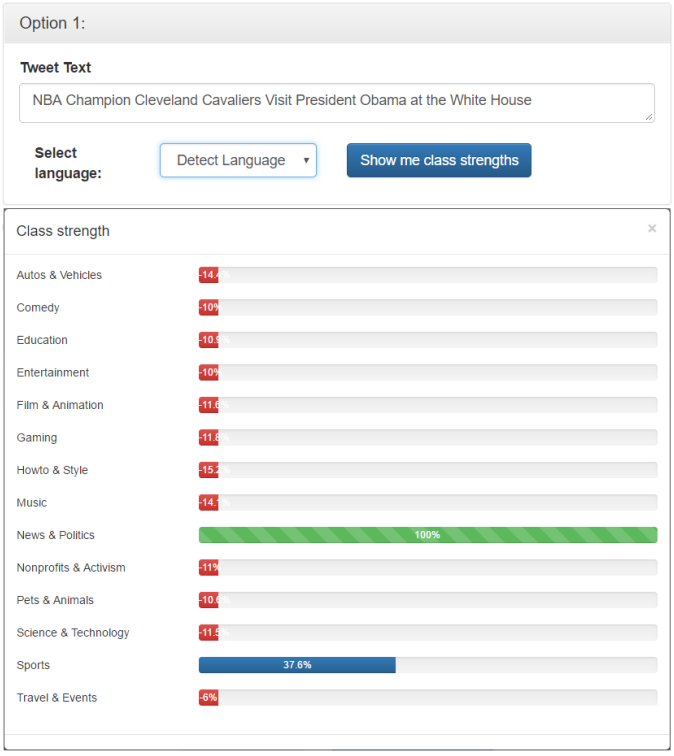

Fig. 2: Example of classification strength of an English tweet using *ClassStrength* online mode

of tweets. Otherwise, the language detector module would detect the language of the tweets in the file and classifies them accordingly.

The input file should contain a tweet in each line with the following formant: Tweet_ID <tab> Tweet_text

The output of classification would contain the tweet ID, followed by a list of 14 numbers representing the strength of each class. The strength of each class is a number from -100 to 100, where negative values mean that these categories are not applicable to this tweet.

Offline Mode: The offline mode allows users to download a full version of the tool to be used locally on their machine. Some system requirements are essential to get it working as described in the ReadMe file provided with the tool package.

V. Conclusion and Future Versions

In this paper we offer our tweet classification tool *ClassStrength*, which is a tool for classifying social text in general and tweets in specific into 14 general-purpose categories. The tool currently works with 5 different languages, namely English, Arabic, French, German, and Russian. The tool uses a state-of-the-art method for collecting training data to train the classifiers. The performance of the tool was reported to be highly satisfactory to users. An online mode of the tool webpage is available for users to test the performance of the tool.

For future versions of *ClassStrength*, we look forward to integrating additional languages, hopefully covering most of

the languages on Twitter. In addition, we will keep updating our classification models by using fresh training data everywhile to keep our classification as accurate as possible. Another version of our classifier is planned to be developed to automatically adapt the classification periodically, since some of the categories drifts over time, especially for topics such as politics and movies.

REFERENCES

[1] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. 2011.

[2] K. Darwish and W. Magdy. *Arabic information retrieval*. Now Publishers, 2014.

[3] D. Irani, S. Webb, C. Pu, and K. Li. Study of trend-stuffing on twitter through text classification. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.

[4] S. Kinsella, A. Passant, and J. G. Breslin. Topic classification in social media using metadata from hyperlinked objects. In *Advances in Information Retrieval*, pages 201–206. Springer, 2011.

[5] A. Kothari, W. Magdy, K. Darwish, A. Mourad, and A. Taei. Detecting comments on news articles in microblogs. In *AAAI press*, 2013.

[6] W. Magdy, H. Sajjad, T. El-Ganainy, and F. Sebastiani. Bridging social media via distant supervision. *Social Network Analysis and Mining*, 5(1):1–12, 2015.

[7] W. Magdy, H. Sajjad, T. El-Ganainy, and F. Sebastiani. Distant supervision for tweet classification using youtube labels. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[8] D. Quercia, H. Askham, and J. Crowcroft. Tweetlda: supervised topic classification and link prediction in twitter. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 247–250. ACM, 2012.

[9] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766. ACM, 2007.

[10] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, pages 42–51. ACM, 2009.

[11] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842. ACM, 2010.

[12] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.