Harnessing Web Page Directories for Large-Scale Classification of Tweets

Arkaitz Zubiaga[1], Heng Ji[2]
Computer Science Department[1,2], Linguistics Department[2]
Queens College and Graduate Center
City University of New York, New York, NY, USA
arkaitz@zubiaga.org, hengjicuny@gmail.com

ABSTRACT

Classification is paramount for an optimal processing of tweets, albeit performance of classifiers is hindered by the need of large sets of training data to encompass the diversity of contents one can find on Twitter. In this paper, we introduce an inexpensive way of labeling large sets of tweets, which can be easily regenerated or updated when needed. We use human-edited web page directories to infer categories from URLs contained in tweets. By experimenting with a large set of more than 5 million tweets categorized accordingly, we show that our proposed model for tweet classification can achieve 82% in accuracy, performing only 12.2% worse than for web page classification.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering

Keywords

tweets, classification, distant, large-scale

1. INTRODUCTION

Twitter's evergrowing daily flow of millions of tweets includes all kinds of information, coming from different countries, and written in different languages. Classifying tweets into categories is paramount to create narrower streams of data that interested communities could efficiently process. For instance, news media could be interested in mining what is being said in the news segment, while governments could be interested in tracking issues concerning health.

Besides tweets being short and informally written, one of the challenges for classifying tweets by topic is hand-labeling a large training set to learn from. Manually labeled sets of tweets present several issues: (i) they are costly, and thus tend to be small and hardly representative of all the content on Twitter, (ii) it is unaffordable to annotate tweets in a wide variety of languages, and (iii) the annotated set might become obsolete as both Twitter and the vocabulary evolve.

Previous research has studied the use of different features for classification purposes. For instance, [5] use a set of social features to categorize tweets by type, e.g., opinions or deals. Similarly, [6] categorized groups of tweets into types of conversations, e.g., events or memes. While they evaluated with 5,407 and 1,036 hand-labeled tweet instances, respectively, studying with larger and more diverse datasets would strengthen the results, so the classifiers could be further generalized to Twitter's diverse contents. In this paper we present a novel approach for large-scale classification of tweets. We describe an inexpensive approach for automatic generation of labeled sets of tweets. Based on distant supervision, we infer the category for each tweet from the URL it points to. We extract categories for each URL from the Open Directory Project (ODP, http://dmoz.org), a large human-edited web directory. Our approach allows to easily create large collections of labeled tweets, as well as to incrementally update or renew these collections. We experiment our proposal with over 5 million tweets, and compare to web page classification, showing that web page directories can achieve 82% in terms of accuracy.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

2. LEARNING MODEL

We propose a learning model based on distant supervision [4], which takes advantage of external information sources to enhance supervision. Similar techniques have been adapted by [2] to categorize blog posts using Freebase, and [1] to infer sentiment in tweets from hashtags. Here, we propose that tweets can be assumed to be in the same topic as the URL they are pointing to, provided that the tweet is referring to or discussing about the contents of the web page. We set out to use the ODP, a community-driven open access web page directory. This way, we can incrementally build a labeled set of tweets, where each tweet will be labeled with the category defined on the ODP for the URL being pointed.

The ODP is manually maintained continuously. However, since Twitter evolves rapidly, its stream is riddled with new URLs that were not seen before. Thus, most URLs found in tweets are not likely to be on the ODP, and the coverage would be very low. To solve this, we consider the domain or subdomain of URLs, removing the path. This enables to match a higher number of URLs in tweets to (sub)domains previously categorized on the ODP. Thus, a tweet pointing to `nytimes.com/mostpopular/` will be categorized as *News* if the ODP says that `nytimes.com` belongs to *News*.

3. DATA COLLECTION

We collected tweets containing links, so we could categorize tweets from the URLs they pointed to. Given the restricted access for white-listed users to the stream of linked tweets, we tracked tweets containing *'http'* from Twitter's streaming API, and removed the few tweets that mentioned

the word 'http' instead of containing a URL. We collected tweets during two non-consecutive 3-day periods: March 20-22, and April 17-19, 2012. Having two non-consecutive timeframes enables to have the former as training data and the latter as test data, with a certainty that tweets from the test set are unlikely to be part of the same conversations as in the training data, avoiding overfitting. The process above returned 12,876,106 tweets with URLs for the training set, and 12,894,602 for the test set. Since URLs in tweets are usually shortened, we retrieved all the unshortened destination URLs, so we could extract the actual domain each tweet points to. For matching final URLs to their corresponding ODP categories, we defined the following two restrictions: (i) from the list of 16 top-level categories on the ODP, we removed 'Regional' and 'World' since, unlike the other categories, do not represent topics, but a transversal categorization based on countries and languages; we keep the remaining 14 categories, and (ii) a small subset of the domains are categorized in more than one category on the ODP; we exclude these to avoid inconsistencies, and consider those with just one category. Matching our tweet dataset to ODP, we got 2,265,411 tweets for the training set, and 2,851,900 tweets for the test set (which covers 19.9% of the tweets collected). For each tweet, we keep the text of the tweet, the category extracted from ODP, as well as the text within the <title> tag of the associated URL, so we can compare web page classification and tweet classification. The resulting dataset includes tweets in tens of languages, where only 31.1% of the tweets are written in English.

4. EXPERIMENTS

We use *SVM-multiclass* [3] for classification experiments. We do 10-fold cross-validation by using 10% splits (of ∼226k tweets each) of the training data described above. Since the focus of this study is to validate the learning model, we rely on TF-based representation of the bag-of-words for both tweets (with URL removed for representation) and web pages, and leave the comparison with other representations for future work. We show the accuracy results, representing the fraction of correct categorizations.

Table 1 compares the performance of using either web pages or tweets for training and test. While obviously learning from web pages to classify web pages performs best, given the nature of web page directories, classifying tweets from tweets performs only 12.2% worse, achieving 82% performance. Interestingly, the system can be trained from tweets to categorize web pages as accurately, but classifying tweets from web pages performs 34.7% worse, potentially because web pages do not include the informal vocabulary needed to train. Moreover, if we reduce the training set from 226k to a manually affordable subset of 10k tweets (note that this is still larger than the 5k and 1k used in previous research described above), the overall accuracy of learning from and testing on tweets drops to 0.565 (31.4% loss).

	Tweet	**Web**
Tweet	0.824	0.830
Web	0.542	0.938

Table 1: Classification of tweets (rows = train, columns = test).

Table 2 further looks at the performance of learning from and testing on tweets. After computing the cosine similarity of words in tweet texts and words in associated URL titles, we break down the accuracy into ranges of tweet-title similarity. This aims to analyze the extent to which tweets can be better classified when they are similar to the title of the web page (more akin to web page classification), or when they are totally different (e.g., adding user's own comments). There is in fact difference in that high-similarity tweets (.8-1) are classified most accurately, but low-similarity tweets (0-.2) show promising results by losing only 8.9%.

Similarity	**0-.2**	**.2-.4**	**.4-.6**	**.6-.8**	**.8-1**
Accuracy	0.778	0.808	0.799	0.787	0.854

Table 2: Classification of tweets by similarity.

5. CONCLUSION

We have described a novel learning model that, based on distant supervision and inferring categories for tweets from web page directories, enables to build large-scale training data for tweet classification by topic. This method can help accurately classify the diverse contents found in tweets, not only when they resemble web contents. The proposed solution presents an inexpensive way to build large-scale labeled sets of tweets, which can be easily updated, extended, or renewed with fresh data as the vocabulary evolves.

Acknowledgments. This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF- 09-2-0053 (NS-CTA), the U.S. NSF CAREER Award under Grant IIS-0953149, the U.S. NSF EAGER Award under Grant No. IIS-1144111, the U.S. DARPA Deep Exploration and Filtering of Text program and CUNY Junior Faculty Award. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

6. REFERENCES

[1] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

[2] S. D. Husby and D. Barbosa. Topic classification of blog posts using distant supervision. *EACL*, pages 28–36, 2012.

[3] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, pages 137–142, 1998.

[4] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *ACL/CONLL*, pages 1003–1011, 2009.

[5] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in twitter to improve information filtering. In *SIGIR*, pages 841–842, 2010.

[6] A. Zubiaga, D. Spina, V. Fresno, and R. Martínez. Classifying trending topics: a typology of conversation triggers on twitter. In *CIKM*, pages 2461–2464, 2011.