Short Text Classification in Twitter to Improve Information Filtering

Bharath Sriram, David Fuhry,
Engin Demir, Hakan Ferhatosmanoglu
Computer Science and Engineering Department,
Ohio State University, Columbus, OH 43210, USA
{sriram,fuhry,demir,hakan@cse.ohio-state.edu}

Murat Demirbas
Computer Science and Engineering Department,
University at Buffalo, SUNY, NY 14260, USA
demirbas@cse.buffalo.edu

ABSTRACT
In microblogging services such as Twitter, the users may become overwhelmed by the raw data. One solution to this problem is the classification of short text messages. As short texts do not provide sufficient word occurrences, traditional classification methods such as "Bag-Of-Words" have limitations. To address this problem, we propose to use a small set of domain-specific features extracted from the author's profile and text. The proposed approach effectively classifies the text to a predefined set of generic classes such as News, Events, Opinions, Deals, and Private Messages.

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

General Terms
Algorithms, Performance, Experimentation.

Keywords
Short text, classification, Twitter, feature selection.

1. INTRODUCTION
Twitter[1] is a social networking application which allows people to micro-blog about a broad range of topics. It helps users to connect with their *followers*. The *tweets* from users are referred to as micro-blogs because there is a 140 character limit imposed by Twitter for every tweet. This lets the users present any information with only a few words, optionally followed with a link to a more detailed source of information. The goal of our work is to automatically classify incoming tweets into different categories so that users are not overwhelmed by the raw data. This is particularly useful when Twitter is accessed via hand held devices like smart phones.

Existing works on classification of short text messages integrate messages with meta-information from other information sources such as Wikipedia and WordNet [2,3]. Sankaranarayanan et al [6] introduce TweetStand to classify tweets as news and non-news. Automatic text classification and hidden topic extraction [5] approaches perform well when there is meta-information or the context of the short text is extended with knowledge extracted using large collections.

We propose an intuitive approach to determine the class labels and the set of features with a focus on user intentions on Twitter [4] such as daily chatter, conversations, sharing information/URLs, and reporting news. Our approach is more general when compared with the TweetStand. It classifies incoming tweets into categories such as News (N), Events (E), Opinions (O), Deals (D), and Private Messages (PM) based on the author information and features within the tweets. Experimental results show that classification accuracy is high even without meta-information and the proposed approach outperforms the traditional "Bag-Of-Words" strategy.

Empirical results show that the authorship plays a crucial role in classification. Authors generally adhere to a specific tweeting pattern i.e., a majority of tweets from the same author tend to be within a limited set of categories.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

2. FEATURE SELECTION
Selecting a subset of relevant features for building robust learning models[2] is another research problem. Hence we used a greedy strategy to select the feature set, which generally follows the definitions of classes. We extracted 8 features (8F) which consist of one nominal (author) and seven binary features (presence of shortening of words and slangs, time-event phrases, opinioned words, emphasis on words, currency and percentage signs, "*@username*" at the beginning of the tweet, and "*@username*" within the tweet). In the classification step, the learning model trains itself using these features. Here we discuss how these features may represent certain classes.

Categorization of tweets into the selected classes requires the knowledge of the source of information. Hence, we selected the authorship information as our primary feature. Corporate tweeters generally have different motivations than personal tweeters. While the former generally publish news in a clear form, the latter instead frequently express themselves by using slang words, shortenings and emotions. Thus, a feature for discriminating news may be the absence of shortenings, emotions, and slang words. This feature can be further used to differentiate the personal tweeters from corporate tweeters.

If we define an event as "something that happens at a given place and time", the presence of participant, place, and time information could determine the existence of an event in the text. Hence, we extracted the date/time information and time-event phrases which are collected from a set of tweets based on general observation of users and set the presence of them as a feature. Participant information is also captured via the presence of the '@' character followed by a username within tweets.

Presence of opinions is determined by a lookup in a wordlist which consist of about 3000 opinionated words obtained from the Web.

[1] http://www.twitter.com

[2] http://en.wikipedia.org/wiki/Feature_selection

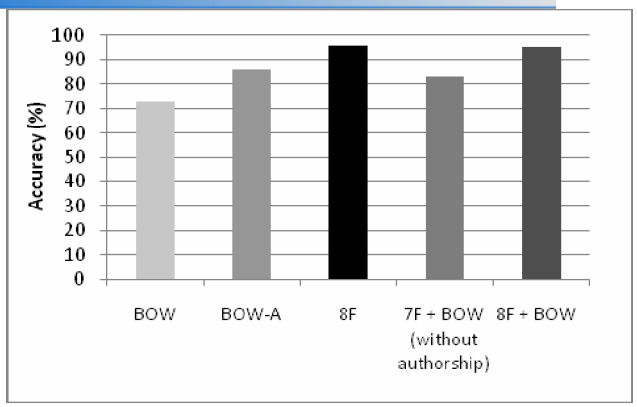

Figure 1. Overall accuracies.

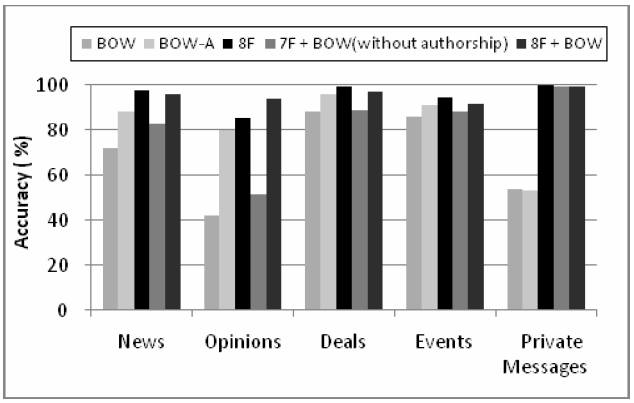

Figure 2. Accuracies for individual classes.

We also capture the emphasis on words based on the usage with uppercase letters. Another way to detect the emphasis is the usage of repeating characters in a word (e.g., "veeery").

The keyword "deal" and special characters within the text such as currency and percentage signs are good features to capture the context of deals.

Twitter lets the users send private messages to other users by using the '@' character followed by a username at beginning of the tweet. Hence, private messages are captured by the usage "@username" at the beginning of tweets.

3. EXPERIMENTAL RESULTS

3.1 Experimental Setup

We downloaded a collection of recent tweets from random users and eliminated the ones not in English, with too few words (threshold set as three), with too few words apart from greeting words, with just a URL, and with too few words apart from URL. Our final collection is composed of 5407 tweets from 684 authors. These tweets were manually labeled with the best matching category (i.e., 2107 N, 625 O, 1100 D, 1057 E, and 518 PM). After removing the stop words, there are 6747 unique words.

Experiments are conducted with the available implementation of Naïve Bayes classifier in WEKA[3] using 5-fold cross validation.

3.2 Performance Evaluation

In Figs 1 and 2, BOW, BOW-A, and 8F refer to Bag-Of-Words, BOW with the author feature, and our approach, respectively. As shown in Fig 1, 8F achieves 32.1% improvement over BOW on the overall accuracy. The author feature is found to be very discriminative in our dataset. BOW-A achieves 18.3% improvement over BOW, and even 3.7% over 7F+BOW (without authorship) on the overall accuracy.

As shown in Fig 2, 8F performs consistently better for all classes. It may be used with BOW to have a better accuracy with an additional time cost of initial training. 8F achieves 35.2%, 103.4%, 12.2%, 9.9%, and 87.0% improvements over BOW for N, O, D, E, and PM, respectively. In BOW, misclassified tweets are mainly between N and PM (383), N and O (407), whereas in 8F, they are mainly between N and O (104). We attribute this to the fact that tweets in N may also be opinionated. We believe that multi-label classification would resolve this issue to a certain extent.

[3] http://www.cs.waikato.ac.nz/ml/weka/

The times taken to build the training models are 37.2 and 0.8 sec for BOW and 8F, respectively. The ratio between these timings will be larger with larger collections as the number of words (features in BOW) will increase, while the feature count in 8F stays fixed.

4. CONCLUSION

We have proposed an approach to classify tweets into general but important categories by using the author information and features within the tweets. With such a system, users can subscribe to or view only certain types of tweets based on their interest.

Experimental results show that BOW approach performs decently but 8F performs significantly better with this set of generic classes. With the usage of a small set of discriminative features, our approach provides a baseline to classify new tweets online with a better accuracy. However, noisier data may degrade the performance of the proposed approach; hence noise removal techniques are necessary in such cases.

We are currently working on incremental classification models to update the set of categories and features dynamically using user's feedback. As a future work, we are planning to support similarity search within our classes supplemented with semantic information gathered from URL information [1] in the tweets. We believe that this will result in higher precision and be especially useful when Twitter is accessed on hand-held devices where performance and accuracy are the major concerns.

5. REFERENCES

[1] Altingovde, I.S., Demir, E., Can, F., and Ulusoy, O. Site-based dynamic pruning for query processing in search engines. In Proc. SIGIR (Singapore, July 2008), 861-862.

[2] Banerjee, S., Ramanthan, K., and Gupta, A. Clustering short text using Wikipedia. In Proc. SIGIR (Amsterdam, The Netherlands, July 2007), 787-788.

[3] Hu, X., Sun, N., Zhang, C., and Chua, T.-S. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In Proc. CIKM (Hong Kong, China, Nov. 2009), 919-928.

[4] Java, A., Song, X., Finin, T., and Tseng, B. 2007. Why we twitter: understanding microblogging usage and communities. In *Procs* WebKDD/SNA-KDD '07 (San Jose, California, August, 2007), 56-65.

[5] Phan, X.-H., Nguyen, L.-M., and Horiguchi, S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proc. WWW (Beijing, China, Apr. 2008), 91-100.

[6] Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, and M. D., Sperling, J. TwitterStand: news in tweets. In Proc. ACM GIS'09 (Seattle, Washington, Nov. 2009), 42-51.