Using External Information for Classifying Tweets

Josh Weissbock
School of Electrical Engineering and Computer Science
University of Ottawa,
Ottawa, ON, Canada
e-mail: jweis035@uottawa.ca

Ahmed A. A. Esmin
Department of Computer Science
Federal University of Lavras,
MG, Brazil
e-mail: ahmed@dcc.ufla.br

Diana Inkpen
School of Electrical Engineering and Computer Science,
University of Ottawa,
Ottawa, ON, Canada
e-mail: diana.inkpen@uottawa.ca

Abstract— **Automatic classification of texts by topic is a well-studied problem. Nonetheless, classifying twitter messages by topic is difficult because the messages are short and the features space for classification is very sparse. We propose a method to enhance the text of the messages that contain links with external information such as the title of the web pages, and with the most frequent terms from these web pages. We show that the results of the classification improve substantially when adding this external information.**

Keywords- Twitter Messages, Automatic Text Classification, Machine Learning.

I. INTRODUCTION

Twitter is a microblog service where users post messages ("tweets") of no more than 140 characters. With over 500 million active users[1], Twitter represents one of the largest and most dynamic real datasets of user-generated and distributed content. Twitter became an important forum for peer interaction and represents a dynamic and interactive tool used by many users, including media and the news companies. The tweets can be used to inform and sometimes express opinions and sentiments about different topics. With these data, companies have the opportunity to reach their users and examine the feedback of customers about products and services. By default, twitter accounts are made public, such that if any one knows your account name they are able to view all of the tweets that you have previously published. Tweets provide additional challenges compared to other text; they are short and include informal / colloquial / abbreviated language.

Although automatic text-classification by topic is a well-studied problem, classifying tweets is difficult because the messages are short and the features space for classification is very sparse. In this paper, we address the task of classification of tweets by using external information from the web pages, given by the URLs in the tweets, such as the title of the webpage and the most frequent terms. Given a set of tweets, we train a classifier on a set of tweets annotated with their categories. Then, the classifier can be used to classify any new tweets according to the same set of categories. Our hypothesis is that by expanding on these tweets we are able to improve the classification rate.

The remainder of this paper is organized as follows. Section 2 briefly presents related work. Section 3 gives an overview of the dataset and the feature sets used in out experiments. In Section 4 we describe the proposed approach and the experimental results. Finally, in Section 5 we present the conclusions and discuss future work.

II. RELATED WORK

Similar work of classification on the web has been conducted before, but not in the same manner or on the same set of texts as we used. By expanding on metadata in hyperlinks (such as on the webpages such as Amazon.com and Youtube.com), Kinsella et al. [1] built a classifier of social media (blogs, message boards, etc.) that was able to achieve an F-Score of 84% to 90%. Pennacchiotti and Popescu [2] used machine learning techniques to construct user profiles as well as the classification of users and their affinity for particular businesses. They showed that machine learning techniques are quite robust across a wide variety of classification problems on the web. Genc et al. [3] expanded upon tweets by matching them to their appropriate Wikipedia page and compared the distances with other tweets. This technique was much more successful than using string edit distance and latent semantic analysis. Xiaoguang and Davison [4] were able to improve classification of web pages by expanding upon the features in neighbouring pages. They found no value in anchor text and unexpected values in title pages. Jiang et al. [5] were able to improve the accuracy of semantic classification of tweets by over 7% by incorporating target-dependent features and looking at related tweets. Many methods of classification of tweets and expanding upon them have been proposed (in the ways mentioned above) and the majority showed that enhancing the tweets with external information is a good idea. To the best of our knowledge, there are no works that improved the classification rate by employing

[1] http://www.mediabistro.com/alltwitter/500-million-registered-users_b18842

methods that enhance the text of the twitter messages that contain links with external information such as the title of the web pages and the most frequent terms from these web pages.

III. The Data Source

We collected a corpus of 9621 training tweets from 6 different CNN twitter accounts using the twitter API and a custom made python script to scrape and save the data. As CNN tweets their news stories to their respective twitter accounts, the tweets were already labeled with categories. The six CNN accounts that we chose were Politics (@CNNPolitics, 1603 tweets), Money (@CNNMoney, 1601 tweets), Entertainment (@CNNShowBiz, 1601 tweets), Sports (@CNNSports, 1604 tweets), Technology (@CNNTech, 1603 tweets) and Travel (@CNNTravel, 1608 tweets). In this way, our training data was conveniently annotated with categories. The distribution of the training data is balanced over the 6 classes, around 16% for each class. Therefore a random classifier would achieve only about 16% accuracy.

A second corpus of 3621 additional tweets was collected from these six twitter accounts, to be used as test data, also with a balanced distribution over the 6 classes (16% each). Their categories are used only for evaluation purposes, to be used for testing the classifiers built on the training corpus. Both corpora are available from the authors upon request.

All usernames were removed from the all the collected tweets. These are the words that begin with an @. For example, "'@BarackObama visited Canada today', cnnpolitics" became "'visited Canada today', cnnpolitics".

IV. The Proposed Approach and the Experimental Results

Initially, we use standard machine learning algorithms from Weka [6] and simple features to train classifiers. We experimented with Support Vector Machines (SVM) (the libSVM implementation) because this algorithm is known to obtain good results on many classifications tasks, with Naïve Bayes because it is known to work well on text classification, and with Decision Trees (DT) (J48 in Weka) because the model that is learnt is human-readable.

For a first experiment, we used the words in the training corpus as features. This is called a bag-of-words representation. We eliminated the stop words from the set of features by using the Python NLTK English stop words resource.

For a more sophisticated method, we prepared an enhanced feature representation. We expanded the corpus of tweets, the first time to get all of the webpage titles for the tweets that had URLs in them. In our dataset, 7314 of the 9621 tweets had URLS to expand, that is 78%. The titles were concatenated to the end of the corresponding tweet. The corpus was expanded a second time to include the top 10 most frequent words from each external webpage. Retrieving these features was done by a Python script that opened every URL, counted the number of instances of every word and then removed all words that are in the Python NLTK 2.0 list of stop words. The top ten remaining words were concatenated to the original tweet with the title. An example of a tweet before and after expansion can be seen in Table I. If no URL was found in the tweet, then the tweet was not expanded. The expanded corpus allows for an enhanced feature representation since now we have more representative features for each topic, from the external web pages. The expansion is likely to lead to better classification results, because the initial representation is limited due to the shortness of the twitter messages.

TABLE I. Example of tweet before and after expansion

Original Tweet	The French and Greek election results are being viewed as a smackdown on austerity. http://t.co/v80GAZSU
Enhanced Tweet	The French and Greek election results are being viewed as a smackdown on austerity. http://t.co/v80GAZSU French and Greek elections: Lessons for U.S. fiscal austerity states united said hollande eurozone plan dont economy

TABLE II. Results of different classifiers

Experiment	**Classifier Model**	Accuracy (%)		
		SVM	**Naïve Bayes**	**Decision Trees**
10-fold cross-validation on training data	OriginalTweets	73.33	71.94	75.79
	EnhancedTweets	**88.27**	79.56	80.68
Test Data	OriginalTweets	39.33	51.28	51.03
	EnhancedTweets	61.05	**62.57**	54.54

By comparing the results of the EnhancedTweets approach with the OriginalTweets classification in Table II, we can see that the accuracy of the classification increased. In the 10-fold cross-validation configuration, the precision of the SVM increased by approximately 14.94 percentage points over the results on the original data model. We looked at the results of cross validation in order to choose the best classifier, which was SVM for our task with an

accuracy of 88.27%. This is why later, in Tables III and IV, we report results for each class using only SVM.

As seen in Table II, on the test data, the accuracy increased by 21.72 percentage points using an SVM model trained on the entire training data set. The accuracy of the other classifiers (Naïve Bayes and Decision Trees) also improved when using the EnhancedTweets approach. Additionally, by looking at the precision, recall and F-measure for each class for both the OriginalTweets in Table III and the EnhancedTweets in Table IV, we see that the F-measures have improved for all the classes (except for Money), while precision and recall improve in 8 of 12 cases. For easier comparison, we summarize the results in Figure 1. Similar results can be seen in Tables V and VI where 10-fold cross-validation was used.

We can also see in Table II, on the train data, the accuracy increased by 14.94 percentage points using an SVM model trained on the entire training data set by using "10-fold cross validation". Figures 1, 2 and 3 show the comparison graphs of the F-measure, Precision and Recall of each class for the two datasets using 10-fold cross validation.

TABLE III. DETAILED RESULTS BY CLASS FOR SVM – TEST DATA (ORIGINALTWEETS MODEL)

Class/Measure	**OriginalTweets model (Test Data)**		
	Precision	*Recall*	*F-measure*
money	0.797	0.457	0.581
politics	0.899	0.193	0.318
showbiz	0.230	0.992	0.373
sportnews	0.982	0.355	0.522
tech	0.759	0.309	0.440
travel	0.412	0.065	0.112
Macro-Average	0.679	0.393	0.389

TABLE IV. DETAILED RESULTS BY CLASS FOR SVM – TEST DATA (ENHANCEDTWEETS MODEL)

Class/Measure	**EnhancedTweets model (Test Data)**		
	Precision	*Recall*	*F-measure*
money	0.929	0.217	0.352
politics	0.793	0.676	0.730
showbiz	0.402	0.962	0.567
sportnews	0.989	0.601	0.748
tech	0.556	0.785	0.651
travel	0.831	0.427	0.564
Macro-Average	0.750	0.611	0.602

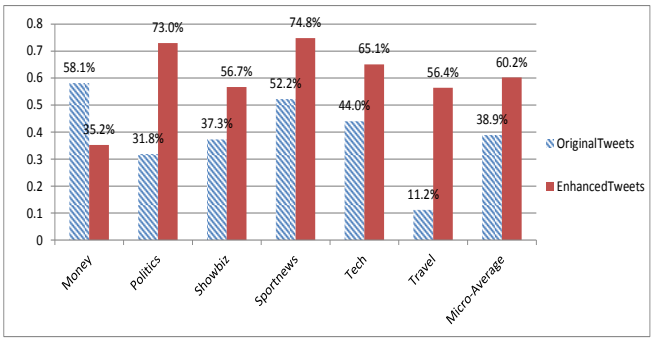

Figure 1. Comparison of F-measure values on the Test Data.

TABLE V. DETAILED RESULTS BY CLASS FOR SVM – TRAINING DATA, 10-FOLD CROSS-VALIDATION (ORIGINAL TWEETS MODEL)

Class/Measure	**OriginalTweets model**		
	Precision	*Recall*	*F-measure*
money	0.845	0.618	0.714
Politics	0.89	0.773	0.827
showbiz	0.439	0.974	0.605
sportnews	0.952	0.868	0.908
tech	0.908	0.631	0.744
travel	0.932	0.537	0.681

TABLE VI. DETAILED RESULTS BY CLASS FOR SVM – TRAINING DATA 10-FOLDS (ENHANCEDTWEETS MODEL)

Class/Measure	**EnhancedTweets model**		
	Precision	*Recall*	*F-measure*
money	0.942	0.798	**0.864**
politics	0.835	0.903	**0.868**
showbiz	0.794	0.96	**0.869**
sportnews	0.99	0.908	**0.948**
tech	0.838	0.876	**0.856**
travel	0.944	0.851	**0.895**

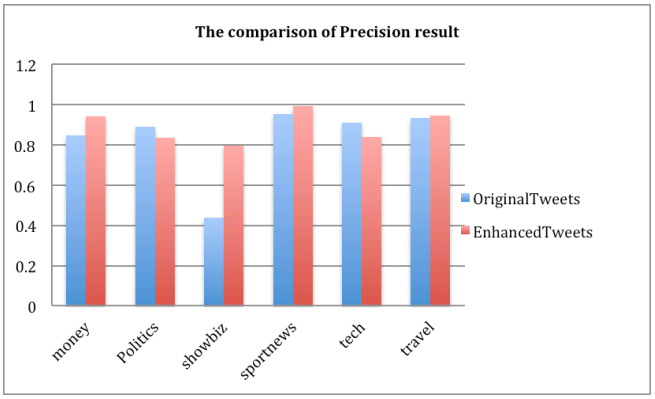

Figure 2. The comparison of Precison results

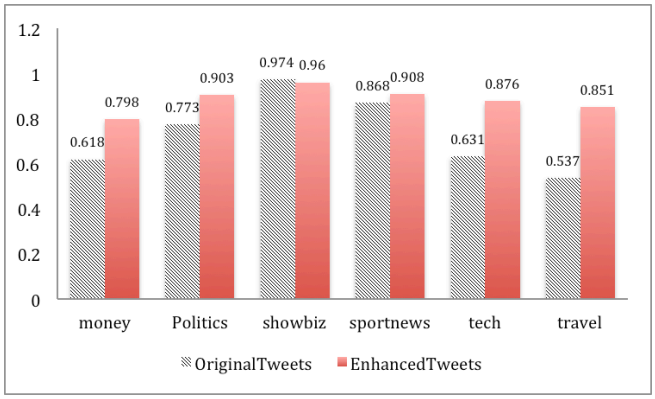

Figure 3. The comparison of Recall results

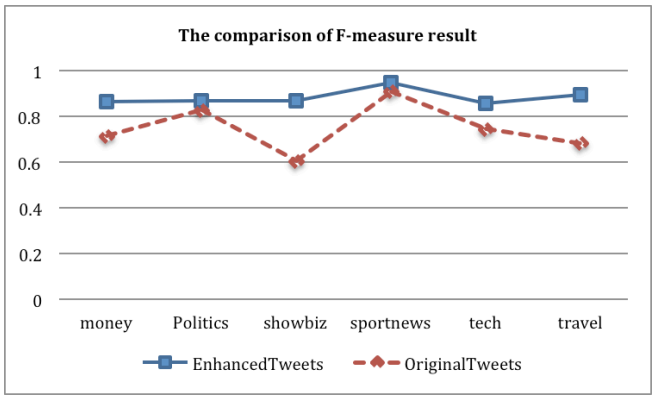

Figure 4. The comparison of F-measure results

The reason for expanding tweets with the top 10 features from the website can be seen in Table VII. By expanding tweets with the top N = 5, 10, 20, or 30 features from each page and running the classifiers, the best results came from using SVM and the top 10 features.

TABLE VII. EXPANSION OF TWEETS WITH N FEATURES FROM EXTERNAL WEBPAGES

N =	5	10	20	30
DT	81.76%	80.68%	83.33%	83.72%
NaiveBayes	77.49%	79.56%	81.43%	80.81%
SVM	85.83%	**88.26%**	87.95%	86.25%

V. CONCLUSION AND FUTURE WORK

We looked at the classification of twitter tweets using machine language algorithms. Cable News Network (CNN) maintains a number of twitter accounts where they tweet their latest news, these accounts are already pre-labeled and most of the tweets follow under that specific category. There are CNN accounts for Entertainment (@CNNShowbiz), Politics (@CNNPolitics), Money (@CNNMoney) etc. Twitter makes these accounts and all their tweets publicly available which we were able to collect and classify. Our initial classification of these tweets were successful only 73% of the time. By expanding upon the web data that is in the URLs of the tweets we were able to add more information. We expanded the tweets to add the web page title, as well as the top ten features on the page (minus the stop words, as mentioned). This expansion led to a successful classification rate of over 88%, an increase of 14.94 percentage points on cross-validation experiments. On the separate test data, we achieved an increase in accuracy from 39.33% to 61.05%.

For further analysis, we looked at the confusion matrix to see which tweets were being misclassified. Many tweets were misclassified as belonging to the showbiz account. This could be because the @CNNShowbiz often retweets (re-publishes) a lot of the posts from the other accounts, rather than just posting Entertainment issues. In future work, we could either remove this account as it is not the best example to use, or use alternative categories such as @CNNHealth. The accounts that are often retweeting other categories may be a cause of the difference of variations between the precision and recall between the different classes. Additionally, we could not look at re-tweets, but only at tweets generated by that account. As tweets posted in quick succession are likely to be related, attempting to join tweets that have been posted within some short time frame would be an alternate suggestion to improve tweet classification.

This success of classification leads us to believe that the expansion of tweets is successful and we can assist classification algorithms by finding more ways to expand the tweets. In this direction, future work could look at who is re-tweeting, relate tweets to Wikipedia articles, use the metadata in the URLs or look at the neighbouring pages of websites.

ACKNOWLEDGMENT

We would like to thank FAPEMIG (Brazil) and NSERC (Canada) for partial financial support. The authors also thank the anonymous reviewers for useful remarks and suggestions.

REFERENCES

[1] Kinsella, S., Passant, A., & Breslin, J. G. (2011). Topic Classification in Social Media Using Metadata from Hyperlinked Objects. (P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, & V. Mudoch, Eds.)*Advances in Information Retrieval Proceedings ECIR 2011*, *1380*(January), 201-206. Springer. Retrieved from http://www.springerlink.com/index/V7N118133XR22264.pdf

[2] Pennacchiotti, M. (2011). Democrats , Republicans and Starbucks Afficionados : User Classification in Twitter. *Statistics*, 430-438. ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2020477

[3] Genc, Y., Sakamoto, Y., & Nickerson, J. V. (2011). Discovering Context: Classifying Tweets through a Semantic Transform based on Wikipedia. *HCI*, 484-492. Springer-Verlag. Retrieved from http://cog.mgnt.stevens-tech.edu/~yasu/papers/hciiclassification2011.pdf

[4] Qi, X., & Davison, B. D. (2008). Classifiers Without Borders : Incorporating Fielded Text From Neighboring Web Pages. *Text*, 643-650. ACM Press. Retrieved from http://portal.acm.org/citation.cfm?id=1390443

[5] Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent Twitter Sentiment Classification. *Computational Linguistics*, *1*(June 19-24, 2011), 151-160. Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P11-1016

[6] Witten, I.H., Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco.