Tweet classification using semantic word-embedding with logistic regression

Muhammad Rafi
National University of Computer and Emerging Sciences, Karachi.
muhammad.rafi@nu.edu.pk

Saeed Ahmed
National University of Computer and Emerging Sciences, Karachi.
k142142@nu.edu.pk

Fawwad Ahmed
National University of Computer and Emerging Sciences, Karachi.
k142051@nu.edu.pk

Fawzan Ahmed
National University of Computer and Emerging Sciences, Karachi.
k142330@nu.edu.pk

ABSTRACT
The paper presents a text classification approach for classifying tweets into two classes: availability/ need, based on the content of the tweets. The approach uses a language model for classification based on word-embedding of fixed length to get the semantic relationship among words. The approach uses logistic regression for actual classification. The logistic regression measures the relationship between the categorical dependent variable (tweet label) and a fixed length words embedding of the tweet-content(words), by estimating the probabilities of tweets produced by embedding words. The regression function is estimated by maximum likelihood estimation of composition of tweets by these embedding words. The approach produced 84% accurate classification for the two classes on the training set provided for shared task on "Information Retrieval from Microblogs during Disasters (IRMiDis)". as a part of, The 9th meeting of Forum for Information Retrieval Evaluation (FIRE 2017).

Keywords
Text classification, word embedding, logistic regression

1. INTRODUCTION
The proliferation of social media messaging sites enable users to get real-time information in case of disaster events. The effective management of disaster relief operations very much dependent on identifying needs and availability of various resources like food, medicine and shelters etc. Considering a large number of tweets during such event, demands to have an automatic way to sort them out and effectively utilizing this information is growing concern now. Twitter is a very popular microblogging platform and generates about 200 million tweets per day. Users post short text of 140 characters of length for communication and this text can be viewed by user's followers and can be searched via tweeter's search. The text classification for such short, often multi lingual text is very challenging and posed a lot of problems [1]. A very challenging problem is to classify the tweets by analyzing the content in a scenario of a disaster like flood or earthquakes in term of whether the tweet is about the availability of a resource for relief or there is some need of a particular resource at some place. The shared task on "Information Retrieval from Microblogs during Disasters (IRMiDis)" [2]. as a part of, the 9th meeting of Forum for Information Retrieval Evaluation (FIRE 2017) is about same. We proposed a language modeling based approach to this classification task. We learned the word embedding of the term present in the tweets using neural network models, the idea is very similar to work in [4]. These words embedding create a semantic vector space. Each tweet is later represented by a projected vector space of word-embedded fixed length vectors, thus producing an automatic semantic feature [3]. We used logistic regression for the task of classification [5].

FIRE'17-irdimis, December 8-10, 2017, Bangalore, India.

DOI: http://dx.doi.org/10.1145/12345.67890

2. METHODOLOGY
Our approach is divided into three phases. In phase one, we preprocessed the tweets. The dataset is preprocessed by performing parsing, stop-word removal and stemming using Porter algorithm. We have selected the textual features using term frequency inverse document frequency (tf*idf) weighting scheme. For multi lingual text, we simply use translation mechanism, all non -English tweets are translated using Google Translator API into English equivalent text. We also filter out the URLs and Emojis text from the tweets. In the second phase, we have created a fixed length word – embedding of each terms from the tweet selected based on tf*idf scores. This phase adds semantic knowledge to the given instance of the tweet using a neural network model for word-embedding. This is particularly meaningful for short text tweets and resolved the issues of sparsity, contextualization and representation. In the final phase, we trained a logistic regression based classifier. The classification process through logistic regression measure the relationship between the categorical dependent variable (tweet label) and a fixed length words embedding of the tweet-content(words), by estimating the probabilities of tweets produced by embedding words. The regression function is estimated by maximum likelihood estimation of composition of tweets by these embedding words.

3. RESULTS AND OBSERVATIONS
The training dataset contains 856-tweets from which 665 for availability and 191 for needs. We decided to build model from a balance dataset took sampled a subset of data for training about 200 of availability and 191 needs. We split the data into two sets on for training and testing the model. The testing of the model was performed on 192. The accuracy of the model 74% from which 85 availabilities and 65 needs tweets were identified correctly. In Table 1, we present the result evaluation of our training set. On average MAP for the training set is 0.1499 The testing set

comprises of 47K tweets. We labeled the tweets with our model. The results for the test set is presented in Table 2. The average MAP is 0.0047 We observed that the output file that we submitted for submitted run01, have missed 40298 and have missed 6702 tweets because we were not able to process these tweets because of emoji's and URL text. Hence our MAP values for need tweets at precision@100 is coming to 0.

Availability-Tweets Evaluation		
Precision@100	Recall@1000	MAP
0.2900	0.1430	0.2165
Need-Tweets Evaluation		
Precision@100	Recall@1000	MAP
0.1210	0.0456	0.0833
Average MAP		0.1499

Table 1: Results on training set

Availability-Tweets Evaluation		
Precision@100	Recall@1000	MAP
0.1400	0.0582	0.0082
Need-Tweets Evaluation		
Precision@100	Recall@1000	MAP
0.0000	0.0375	0.0011
Average MAP		0.0047

Table 2: Results on test set

Some examples from the classification task are presented in Table 3 and Table 4.

4. CONCLUSION AND FUTURE WORK

We propose a simple, enrichment based, scalable approach for classification of short tweet text into two classes availability/need. It is worth mentioning that our approach complements the research on enriching short text representation with word embedding based semantic vectors. The proposed approach has a great potential to achieve much better results with more research on (i) term weighting scheme or smoothing (ii) feature selection and (iii) classification methods. Although our initial investigation and experiments are not able to produced exceptional results, we are confident that there are several direction of improvement on our results.

5. ACKNOWLEDGMENTS

Our thanks to IRMiDis Track organizer for providing us an opportunity to work on this interesting problem. We also like to thank computer science department of NUCES FAST Karachi campus.

6. REFERENCES

[1] R. Batool, A. M. Khattak, J. Maqbool and S. Lee, "Precise tweet classification and sentiment analysis," 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS), Niigata, Japan 2013

[2] M. Basu, S. Ghosh, K. Ghosh and M. Choudhury. Overview of the FIRE 2017 track: Information Retrieval from Microblogs during Disasters (IRMiDis). In Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, December 8-10, 2017, CEUR Workshop Proceedings. CEUR-WS.org, 2017.

[3] Vo, Duy-Tin, and Yue Zhang. "Target-Dependent Twitter Sentiment Classification with Rich Automatic Features." IJCAI. 2015.

[4] Tang, Duyu, et al. "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification." ACL (1). 2014.

[5] Genkin, A., Lewis, D. D., & Madigan, D. Large-scale Bayesian logistic regression for text categorization. Technometrics, 49(3), 291-304. 2007.

Availability Tweets Example from the results		
Tweet-ID	**Text**	**Classifier**
592723044302528512	We all are with Nepal at this time of tragedy	0.793487
594215027038494723	sending items to the earthquakw victims We have some mask that we need to send to a company in The earthquake	0.842821

Table 3: Examples of tweets from Availability category

Need Tweets Example from the results		
Tweet-ID	**Text**	**Classifier**
595022733156618240	#Nepal #news Still Needs Five Lacs: The Cooperative Development Ministry distributes about 5 million	0.907212
592955066622939136	I disagree with VHP leaders The world knows Rahul Gandhi is capable of nothing let alone earthquake	0.774131

Table 4: Examples of tweets from needs category