

BITS_PILANI@IMRiDis-FIRE 2017:Information Retrieval from Microblogs during Disasters

Arka Talukdar¹

Rupal Bhargava²

Yashvardhan Sharma³

WiSoc Lab, Department of Computer Science

Birla Institute of Technology and Science, Pilani Campus

Pilani-333031

{f2015112¹, rupal.bhargava², yash³} @pilani.bits-pilani.ac.in

ABSTRACT

Microblogging sites like Twitter are increasingly being used for aiding relief operations during disaster events. In such situations, identifying actionable information like needs and availabilities of various types of resources is critical for effective coordination of post disaster relief operations. However, such critical information is usually submerged within a lot of conversational content, such as sympathy for the victims of the disaster. Hence, automated IR techniques are needed to find and process such information. In this paper, we utilize word vector embeddings along with fastText sentence classification algorithm to perform the task of classification of tweets posted during natural disasters.

CCS CONCEPTS

- **Information Retrieval** → Clustering and Classification;

KEYWORDS

Word embedding, sentence classification, fastText, twitter, multilingual text classification

1 INTRODUCTION

This paper describes our approach for the Microblog Track in FIRE 2017.[1] Microblogging sites like Twitter are important sources of real-time information, and thus can be utilized for extracting significant information at times of disasters such as floods, earthquakes, cyclones, etc. The aim of the Microblog track at FIRE 2017 was to develop IR systems to retrieve important information from microblogs posted at the time of disasters. The task involved identifying tweets to develop automatic methodologies for identifying need-tweets and availability-tweets.

Two classes that were to be identified were defined as:

- (1) Need-tweets: Tweets which inform about the need or requirement of some specific resource such as food, water, medical aid, shelter, mobile or Internet connectivity, etc.

(2) Availability-tweets: Tweets which inform about the availability of some specific resources. This class includes both tweets which inform about potential availability, such as resources being transported or dispatched to the disaster-struck area.

We used word embeddings to represent tweets and then fastText[3] classification algorithm to classify the tweet to its appropriate category. Our system has performed considerably well given its robustness and low resource utilization.

2 BACKGROUND / RELATED WORK

Classification of tweets has been tackled in many shapes and forms over the years. Overtime, we've seen a shift from one-hot vectors representing words to more dense vectors based on word embeddings. Yang et.al, for example, show how leveraging word embeddings can improve classification of tweets to predict election results [7]. Crisis response, in particular, has been tackled leveraging twitter data as well. Imran et al, focuses on building a strong word2vec model based on crisis response tweets and leverages basic linear regression models[2]. Most notably, Zhou et al showcase that c-LSTMs, a hybrid approach between CNNs and LSTMs showed significantly improved results over traditional models when classifying text [8][9]. Using Neural Networks is not the ideal solution due to its high resource requirement. FastText, on the other hand, gives nearly the same performance at a fraction of the resources.

3 DATA

The training data was a collection of about 20,000 tweets posted during the Nepal earthquake in April 2015, along with the associated metadata for each tweet.[1]

The major challenge with the data was that the tweets involved were multilingual and code-mixed. The tweets were in English, Hindi, and Nepali and there were very few training examples in Hindi and Nepali which made it difficult to train. Apart from it, another issue that was faced was the imbalance in class proportions in training data with only a small portion belonging to positive.

Tweets have a stringent word limit, and users often make use of innovative abbreviations which are difficult to handle for retrieval systems. Besides, they are mostly informal and may involve the use of multiple languages in the same tweet (called code mixing), or even multiple scripts in a tweet. It is also difficult to make sense of emoticons, and informal shorthands especially innovative ones made up by users.

4 PROPOSED TECHNIQUE

The problem is formulated as a classification task and the objective is to learn a classifier. The proposed methodology involves a pipelined approach and is divided into four phases:

- Pre-processing of Tweet Corpus
- Creating word-embeddings
- Training classifier
- Calibrate Results with Platt scaling

4.1 Preprocessing

The tweet texts were extracted from associated metadata and were pre-processed in order to ensure uniformity. Pre-processing included removal of emoticon special characters, numbers, hashtags punctuation and words which were not present in Roman and Devnagri script and converting all Roman characters to lowercase.

4.2 Creating Word Embeddings

Created word embeddings using skip-gram and cbow algorithm with the training tweets as the corpus. Best results were obtained with window of size 5 in case of both the algorithms. To create the word-vectors we used google's word2Vec library support.

4.3 Training the model

The fastText classifier is trained on the labeled data and the previously created word embeddings. FastText is optimised and learning rate and other hyper parameters are tuned using grid search.[4] FastText creates sentence vectors from the individual word vectors in the words of the tweets. The fastText algorithm uses this sentence vectors to classify tweets.

4.4 Calibrating the model

The fastText classifier tends to push probabilities to the extremes, such a model is not well calibrated. To calibrate the model and ensure even distribution of probability, Platt scaling is applied.[8] The final scaled probability of tweet was used to rank the tweets in each category.

5. EVALUATION RESULTS

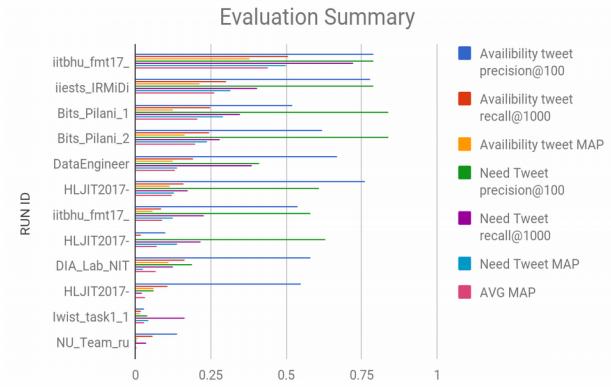
The results of both the runs have been summarized in the following tables:

BITS_PILANI_RUN1: Skip-Gram word embeddings:

| Class | Precision@100 | Recall@1000 | MAP |
|--------------|---------------|-------------|--------|
| Need | 0.84 | 0.3466 | 0.2903 |
| Availability | 0.52 | 0.25 | 0.1244 |

BITS_PILANI_RUN2: CBOW word embeddings:

| Class | Precision@100 | Recall@1000 | MAP |
|--------------|---------------|-------------|--------|
| Need | 0.84 | 0.281 | 0.2362 |
| Availability | 0.62 | 0.2459 | 0.1625 |



Both the runs yielded similar results with skip-gram performing marginally better than CBOW. The runs achieved highest precision@100 among all submissions while it gave reasonable in recall@1000.

6 CONCLUSION

Information available on social media platform, like twitter, during an emergency situation proved to be immensely useful for crisis response and management. However, analyzing large amounts of social media data pose serious challenges to crisis managers, especially under time-critical situations. In this paper, we presented a method that trains very fast and at low resource to effectively monitor social media big crisis data in a timely manner.

The proposed model can be improved significantly if its recall is improved. A major cause of poor recall was the imbalance in the dataset, data augmentation techniques may resolve this issue, this can be extended as future work.

REFERENCES

- [1]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

[2] Imran, Muhammad, Prasenjit Mitra, and Carlos Castillo. "Twitter as a lifeline: Human annotated twitter corpora for NLP of crisis-related messages." arXiv preprint arXiv:1605.05894 (2016).

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.

[4] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

[5] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "AIDR: Artificial intelligence for disaster response," in Proceedings of the companion publication of the 23rd international conference on World wide web companion. International World Wide Web Conferences Steering Committee, 2014, pp. 159–162.

[6] "Tweepy." Tweepy. N.p., n.d. Web. 22 Mar. 2017.

[7] Yang, Xiao, Craig Macdonald, and Iadh Ounis. "Using word embeddings in twitter election classification." arXiv preprint arXiv:1606.07006 (2016).

[8] Hsuan-Tien Lin, Chih-Jen Lin, Ruby C. Weng, A note on Platt's probabilistic outputs for support vector machines, Machine Learning, v.68 n.3, p.267-276, October 2007

[9] CS224N Final Project: Detecting Key Needs in Crisis. Tulsee Doshi (tdoshi), Emma Marriott (emarriott), Jay Patel (jayhp9). March 22, 2017