DataBros@Information Retrieval from Microblogs during Disasters(IRMiDis)

Naveen Kumar
IIIT Kalyani
A10/137, IIIT Kalyani Boys Hostel
Kalyani, West Bengal 741235
naveen.pwn@gmail.com

Mradul Dubey
IIIT Kalyani
A10/137, IIIT Kalyani Boys Hostel
Kalyani, West Bengal 741235
mraduldubey@iiitkalyni.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.[1]

CCS CONCEPTS

•**Data Science →Machine Learning;** *NLP;* Tweet Extraction;

KEYWORDS

Data Mining, NLP, Machine Learning

1 INTRODUCTION

In this track, focus is on two types of tweets:

1.1 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. Note that tweets which do not directly specify the need, but point to scarcity or non-availability of some resources (i.e., a covert expression of the need) are also included in this category. For instance, the tweet "Mobile phones not working" is considered as a need-tweet, since it informs about the need for mobile connectivity.

1.2 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, as well as tweets informing about the actual availability in the disaster-struck area, such as food being distributed, etc. Note that a particular tweet may be both a need-tweet and an availability-tweet if it informs about the need of some specific resource, as well as the availability of some other resource.

The track will have two sub-tasks, as described below:

1.3 Sub-task 1: Identifying need-tweets and availability-tweets

Here the participants need to develop automatic methodologies for identifying need-tweets and availability-tweets. This is mainly a search problem, where relevant microblogs have to be retrieved. However, apart from search, the problem of identifying need-tweets and availability-tweets can also be viewed as a pattern matching problem, or a classification problem (e.g., where tweets are classified into three classes- need-tweets, availability-tweets, and others).

1.4 Sub-task 1: Matching need-tweets and availability-tweets

An availability-tweet is said to match a need-tweet, if the availability-tweet informs about the availability of at least one resource whose need is indicated in the need-tweet. Table 1 shows some examples of need-tweets and matching availability-tweets. In this sub-task, the participants are required to develop methodologies for matching need-tweets with appropriate availability-tweets. Note that an availability-tweet is considered to match a need- tweet even if there is a partial match of the resources, e.g. if the need-tweet mentions about multiple resources and the availability-tweet inform the availability of a subset of these resources. Also, note that a need-tweet and a matching availability-tweet can be in different languages; either or both might be code-switched as well.

2 METHODOLOGIES

2.1 Dataset & Preprocessing

The python code which was provided for us was used to crawl both train data and test data. The twitter data was crawled in json format. It was then converted into a csv file by taking tweet-id, text, and its class as attribute. Classes were assumed as 0 for non-relevant tweets, 1 for need-tweets, and 2 for availability-tweets.

All characters other than alphabets were removed from our tweets and converted them in small letters. Stopwords were also removed and then stemming was done so that similar words with different verb forms could be treated as same. After that, most common words among the tweets were found. They were removed except some selected words which are [medical, need, give, relief, fund, food, donate, aid, water, meal, send, offer, finance, blood]. Also all the retweets and redundant tweets were removed.

2.2 Model

(1) First model was a simple Bag-Of-Words (BOW) model. It selects the features from the tweets as vocabulary and keeps most important features at the top. It gave a good result but it was not enough.

(2) TfidfVectorizer was used to collect the features. It included the unigrams and bigrams. Limit to max features extraction was kept to 6000. After that Recursive Feature Elimination (RFE)[3] was used with LinearSVM as estimator to select 1000 most informative features. The main purpose of SVM-RFE is to compute the ranking weights for all features and sort the features according to weight vectors as the classification basis. SVM-RFE is an iteration process of the backward removal of features. Its steps for feature set selection are shown as follows:
   (a) Use the current dataset to train the classifier.
   (b) Compute the ranking weights for all features.
   (c) Delete the feature with the smallest weight.

   Implement the iteration process until there is only one feature remaining in the dataset; the implementation result provides a list of features in the order of weight. The algorithm will remove the feature with smallest ranking weight, while retaining the feature variables of significant impact. Finally, the feature variables will be listed in the descending order of explanatory difference degree. SVM-RFE's selection of feature sets can be mainly divided into three steps, namely,
   (a) the input of the datasets to be classified,
   (b) calculation of weight of each feature, and
   (c) the deletion of the feature of minimum weight to obtain the ranking of features.

   After this, the classifier used to classify our data was DecisionTreeClassifier[4]. The tweets were ranked according to the probability of it being to that class.

2.3 Validation

Data was split as 80% for training and 20% for testing and our model gave 84% accuracy.

2.4 Task-2

Need-tweets and availability-tweets were available after the classification step. POS (Parts of Speech) tagging was used to remove all words other than Common Nouns from our need-tweets and availability tweets. There is a nice paper on POS tagging with high accuracy[2]. Now for each word in need tweet, that word was searched in availability-tweets. If there is a match, it can be said that availability-tweet matches need-tweet.

3 RESULTS

After the results were declared by the organizers, the following results were obtained:

Task1:

Availability-Tweets Evaluation			Need-Tweets Evaluation			Average MAP
Precision@100	**Recall@1000**	**Map**	**Precision@10**	**Recall@1000**	**Map**	**MAP**
0.7800	0.3031	0.2126	0.7900	0.4052	0.3152	0.2639

Task2:

Precision@5	Recall	F-Score
0.2482	0.3888	0.3030

4 FUTURE WORK

Lemmatization can be used in place of Stemming which will give more accurate context. Also, PCA can be used in place of RFE. Future work can include using spellchecker and correcting it, using wordnet for getting more accurate features and will make our model more-flexible.

Task-2 can be much improved by finding more accurately what is needed and also in which exact location. Need-tweet can be matched with that availability-tweet which can fulfill most of its demands or if the quantity is given of the needed thing, it can be matched to that availability-tweet that has that amount of the needed thing. Also, it will be helpful to match the tweet with that tweet which is more geographically closer. Language barrier can be removed by making the model enable to understand major languages.

REFERENCES

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

[2] C.D.Manning. 2011. Part-of-Speech Tagging from 97Time for Some Linguistics? (2011). https://doi.org/pubs/CICLing2011-manning-tagging.pdf

[3] Lee WM Li RK Jiang B-R Huang M-L, Hung Y-H. 2014. SVM-RFE Based Feature Selection and Taguchi Parameters Optimization for Multiclass SVM Classifier. *The Scientific World Journal* (2014). https://doi.org/pmc/articles/PMC4175386/

[4] Witten D. Hastie-T. Tibshirani R. James, G. 2013. *An Introduction to Statistical Learning*. Springer.