Fully Automatic Approach to Identify Factual or Fact-checkable Tweets

Sarthak Anand[1], Rajat Gupta[2], Rajiv Ratn Shah[3], and Ponnurangam Kumaraguru[3]*

[1] Netaji Subhas Institute of Technology, New Delhi 110078, INDIA
[2] Maharaja Agrasen Institute of Technology, Delhi 110086, INDIA
[3] Indraprastha Institute of Information Technology, Delhi 110020, INDIA
sarthaka.ic@nsit.net.in

Abstract. This paper presents the solution of the team MIDAS of IIIT Delhi for the IRMiDis track in FIRE 2018. We present our solution for the identification of factual or fact-checkable tweets from a dataset that consists of about 50,000 tweets posted during the 2015 Nepal earthquake. We provide a rule based approach for this task and compare it with a semi-supervised approach. After preprocessing steps including tokenization and cleaning, we calculate a factuality score on the basis of number of proper-nouns and quantitative values within a tweet and finally rank them according to the score. Experimental results show that this simple rule based approach provides comparable results in comparison to that of semi-supervised approach.

Keywords: Social media analysis, Unsupervised learning, Information retrieval, Microblogs, Disaster

1 Introduction

Social media usage has considerably increased over the last decade. People often use the social media for various purposes and create a huge amount of user-generated content. In addition to the reporting of news or events social media platforms are increasingly being used for aiding relief operations during various mass emergencies, *e.g.*, during Kerala floods 2018.

However, messages posted on these sites often contain rumors and false information. In such situations, identification of factual or fact-checkable tweets, *i.e.*, tweets that report some relevant and verifiable fact is extremely important for effective coordination of post-disaster relief operations. Additionally, cross verification of such critical information is a practical necessity to ensure the trustworthiness. Considering the scale of these platforms it is not feasible to manually check and verify different user-generated content on time. Since it is very important to reach to a person who is stuck in such emergencies on time,

[*] This work was done when Ponnurangam Kumaraguru was on sabbatical at International Institute of Information Technology, Hyderabad.

automated IR techniques are needed to identify, process and verify the credibility of information from multiple sources.

With this paper we provide one such approach which has shown the best performance in the FIRE challenge 2018 [1] on identifying factual tweets.

2 Related Work

Identifying factual and non factual tweets can be treated as a supervised classification problem. A lot of work have already been done related to supervised based classification [7] [8] [4]. All these works require large amounts of manually labeled dataset.

Despite most works focus on supervised techniques, some works also employed unsupervised techniques as well. For instace, Bjorn Schuller et al. [6] worked on knowledge based approach which does not demand labeled training data. Moreover, Shailesh S. Deshpand et al. [2] proposed a rule based approach for the classification of sentences. They tested it for identifying specific and non specific sentences. They computed several features for each sentence for computing a specificity score for each sentence. Similar to their approach we extract features from sentences such as the number of proper nouns(PROPN) and the number of quantitative values(NUM) and compute a factuality score(*higher score indicates more factual information*). In our approach, we use the factual score for ranking the tweets in order of factual information and use the top k sentences as fact-checkable tweets.

3 Problem and Data Description

Information retrieval from micro-blogs during disasters challenge had 2 subtasks. Sub-task 1 was about, identifying factual or fact-checkable tweets related to Nepal disaster and ranking them on the basis of their factuality scores. Subtask 2 was about, mapping the fact-checkable tweets with appropriate news articles. The submission was categorized into 3 types based on the amount of manual intervention i.e. **Fully automatic**, **Semi automatic**, and **Manual**.

Data Description Dataset for sub-task 1 consists of about 50,000 tweets posted during the 2015 Nepal earthquake. Dataset for sub-task 2 included around 6,000 news articles related to the 2015 Nepal earthquake. Refer [1] for more details.

4 Automatic Methodology

The problem at hand is to use tweets and rank them based on the information they contain. The following sections describe in detail the various steps that have been performed to achieve the results and intuition behind our approach.

1. Pre-processing of tweets, POS tagging and finding proper-nouns and quantitative values, are described in Section 4.2.

2. Finally computing a factuality score based on proper-nouns and quantitative values, is described in Section 4.3.

4.1 Intuition

Similar to the findings of Shailesh S. Deshpande et al. [2], in our study we find that tweets that contain some factual information consists of some name entities like an organization like *UN or NDRF*, or proper noun such as *PM Modi* and quantitative information such as date, time or numbers(*e.g., 5 dead or 5 tonnes*). Based on this study we try to score a tweet on the basis of number of proper nouns and quantitative values which we call as factuality score.

4.2 Data Preprocessing and POS tagging

Since the data given to us is raw, noisy and also prone to more errors, it cannot be directly used for analysis. It is necessary to perform some preprocessing to make the data more suitable so that we can perform POS tagging on the sentences. The following preprocessing steps were performed:

1. **Tokenization:** Tokenization refers to the breaking down of the given text into individual words. We use the Spacy's word tokenizer to perform tokenization of the tweets.
2. **Normalization:** We perform the following steps, very specific to tweets to normalize our corpus:
   - **Stop-words and punctuation removal:** Usually tweets consists of mentions, hash-tags, URLs, punctuation marks and emoji's. They are not useful in determining the amount of information within a tweet and hence are removed from our corpus.

POS tagging In our approach, we have used two major features for computing factuality score, *i.e.*, the number of proper nouns and quantitative values within a tweet. We use spacy's POS tagger for this purpose.

4.3 Computing Factuality Score

Submitted Approach [1] In this approach we compute the number of proper nouns and number of quantitative values within the tweet. For mapping the score to 0 and 1 we divide the number of PROPN and NUM by maximum values achieved in their respective field. Finally, we take average of both these values. The Table 1 shows examples for calculating the factuality score. The underlined words refer to proper-nouns and italicized words refer to numbers. For these examples, note that the maximum values of PROPN and NUM were 17 and 13, respectively. (Shortcomings and suggestions for this approach are described in Section 7)

[1] Github code available at: `https://github.com/isarth/Fire_task_1`

Table 1. Computing factuality score

Tweet	PROPN score	NUM score	Factuality score
Currently working rescue Army CHINA, INDIA, FRANCE, ISREAL, TURKEY, GERMANY, USA, UK, UAE,	13/17= 0.7647	0.0	0.3823
Missing tourists in earthquake include *15* French, *12* Russians,*10* Canadians, *9* Americans and *8* Spanish.	5/17 = 0.2941	5/13=0.3846	0.3393
Quake wake-up call for govt, need better building tech: Eert	0	0	0

5 Semi Automatic Methodology

For comparing our automatic approach with supervised approach. We manually labeled around 1,500 tweets as factual and non-factual and treat the sub-task 1 (refer Section 3) as binary classification problem. The confidence score of the classifier is treated as the factuality score, which is finally used for ranking the tweets. The following section describes in detail various steps that have been performed for the semi-automatic approach.

1. Manually labeling a small set of tweets from the dataset.
2. Pre-processing steps, already described in Section 4.2
3. Training a binary classifier and finally ranking tweets according the confidence score (see Section 5.1 for details).

5.1 Binary Classifier

For classifying tweets as factual and Non-factual, we train both Fasttext [3] cbow and bi-gram models. We split our labeled dataset into two parts training and validation. Table 2 shows the performance of both the classifiers. Finally for ranking tweets in order of factuality, we treat the confidence score of bi-gram classifier as our factuality score.

Table 2. Performance of FastText classifiers

Fast Text classifier	Validation Accuracy
CBOW	0.756
Bi-gram	0.796

6 Result and Analysis

Finally Table 3 compares the results of automatic and semi-automatic approach in the FIRE'18 challenge. Table 4 summarizes the final results of other teams that participated in the FIRE'18 task for automatic submission. We were ranked first in the competition with an NDCG score of 0.6835. The lowest NDCG score achieved in the competition was 0.1271. Table 5 summarizes the final results of other teams that participated in the FIRE'18 task for semi-automatic submission. We were ranked second in that task. For detailed results refer [1].

Table 3. Results show that automatic approach is comparable with semi automatic approach.

Method	P@100	R@100	MAP	NDCG@100	NDCG
Semi Automatic	0.960	0.1148	0.1345	0.6007	0.6899
Automatic	0.880	0.1292	0.1329	0.5649	0.6835

Table 4. Top five automatic submission (our submission: MIDAS)

Teams	P@100	R@100	MAP	NDCG@100	NDCG
MIDAS et al.	***0.8800***	***0.1292***	***0.1329***	***0.5649***	***0.6835***
FASTNU et al.	0.7000	0.0885	0.0801	0.5723	0.6676
UEM et al.	0.6800	0.1427	0.1178	0.5332	0.6396
UEM et al.	0.6400	0.1069	0.0767	0.5237	0.5276
IIT BHU et al.	0.9300	0.1938	0.1568	0.8645	0.4532

Table 5. Final ranking of all teams for semi-automatic (our submission: MIDAS) submission

Teams	P@100	R@100	MAP	NDCG@100	NDCG
DAIICT et al.	0.400	0.2002	0.1471	0.4021	0.7492
MIDAS et al.	***0.9600***	***0.1148***	***0.1345***	***0.6007***	***0.6899***
IIT BHU et al.	0.390	0.0447	0.0401	0.3272	0.620

7 Conclusion and Future Work

We have presented our automatic approach for calculating the factuality score on the basis of number of proper-nouns and quantitative values within a tweet which provided comparable results with semi automatic approach in FIRE'18

Information Retrieval from Micro-blogs during Disasters (IRMiDis) task. The best automatic submission achieved the NDCG score of 0.6835, that made our team stand at first position globally in terms of NDCG score.

On further exploring we find two minor issues in the automatic approach described in Section 4.3 are:

1. Because we are dividing by the maximum value in each field to obtain PROPN and NUM score. The individual scores will not contribute equally for the factuality score.
2. Also, tweets with large number of just quantitative values or proper-nouns will achieve high factuality score.

To overcome the above mentioned issues, we suggest having an upper-bound to the PROPN and NUM values as λ. Hence for computing the individual score we take *min(propn/num, λ)* and finally to map score between 0 and 1 we divide by λ and take the average of both the scores. Futher exploration can be done of finding value of λ. These shortcomings remain, as to be solved as future work.

We also aim to extend the model by making it more efficient by using different techniques we did not explore such as using other features like TFIDF [5] score of words, combined with the ones we already tried. Further knowledge based classification [6] can also be explored .

References

1. Basu, M., Ghosh, S., Ghosh, K.: Overview of the FIRE 2018 track: Information Retrieval from Microblogs during Disasters (IRMiDis). In: Proceedings of FIRE 2018 - Forum for Information Retrieval Evaluation (December 2018)
2. Deshpande, S.S., Palshikar, G.K., Athiappan, G.: Unsupervised approach to sentence classification (2010)
3. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. CoRR **abs/1607.01759** (2016)
4. Lei Shen, J.Z.: Empirical evaluation of rnn architectures on sentence classification task
5. Ramos, J.: Using tf-idf to determine word relevance in document queries
6. Schuller, B., Knaup, T.: Learning and knowledge-based sentiment analysis in movie review key excerpts
7. Wang, S., Manning, C.D.: Baselines and bigrams: Simple, good sentiment and topic classification (2012)
8. Zhang, X., Zhao, J.J., LeCun, Y.: Character-level convolutional networks for text classification. CoRR **abs/1509.01626** (2015)