Extracting Information from Microblogs Posted During Natural Disasters

Santosh Kumar Vishwakarma, Rishabh Singh Chandel, Parul Hora

***Abstract:** Social Networking websites plays an important role in our life. The usage of the above websites is in every domain of our lives and they have increasingly infused itself into daily life. In recent years, the social networking websites such as Twitter, Facebook, are used in natural disasters. Many features have been included in Twitter for fast responses in such kind of unexpected events. This paper is based on the experiments performed on the 2017 Microblog Track provided by Forum of Information Retrieval & Evaluation. The Classification schemes are used with two predefined labels as need and availability. The various pre-processing and natural language processing techniques are applied before the training of the model. The experiments showed that the classification accuracy is increased when the term weight is modified by using the information gain method and using the SVM classifier. This system automatically annotated the FIRE-2015 dataset of microblog track with 97% accuracy.*

***Keywords:** Classification, NLP, FIRE, Information Retrieval, tweets; Natural Disaster; Social Media, disaster monitoring, text classifier Microblogging sites, Twitter, Precision, Recall*

I. INTRODUCTION

There are numerous microblogging sites out of which twitter is one of the most important and widely used platform which people uses to express their views, emotions, comments etc; belonging to wide area like personal, religious, sports, political, entertainment and many more such fields. This enormous data posted by people in raw format can be a source of very useful and meaningful information when it will be mined carefully and intelligently. The FIRE track of Microblogs Disasters aims to train and deploy an information retrieval model that must be capable of extracting information as per the context of the tweets. There are lots of benefits of using the social media applications for the organizations that deals in the disaster response, they usually need the ability and their main task of them include continuously updating the general public with the most precise and important information, when the usual media houses and channels are not available on the spot. They need to be updated by various sources which includes information from the common citizens, affected crowds and they also need to be organizing volunteer efforts and always remains in contact with the populations.

For the natural disaster affected people social media platforms are the quickest methods to let the other people and concerned authorities know about their and other affected people conditions and also a means to get in contact with their friends and family members so that they can be informed and updated about their latest conditions, social media platforms also help them to know about from where they will get the help either from the concerned government authorities or any individual. Meanwhile there is another side of this positive aspect of the coin too that is false and wrong information can be spread too to the concerned authorities and public [1].

The outcome of this wrong and false information can prove to be fatal as mass confusion and lots of chaos can be resulted from both the well-intended as well as the malicious source which finally results in the total waste of time, effort and precious resources which are already in too much demand and valuable in such a critical situation. Also, in many situation same social media platforms are used to make people aware about the pre disaster 5warning, preparedness information, and to make them ready and aware about the detected disaster signals and as an aid of pre disaster warning platform [2].

In this digital revolution, the social media are one of the most important broadcasting mechanisms which facilitate the early warnings mechanism and decision support systems [3].

Twitter is one of the most usable platforms for sharing of information. The twitter users can express and post short text messages, usually known as tweets. These short messages are a medium for disseminating information among the community. During an emergency, the tweets are one of the fastest medium of broadcasting the information.

If we focus on the figures regardless of one’s geographical location the data says that on an average 554.7 million people of the world round the globe are actively using the social media and 58 million tweets are being posted every day [4].

The microblogs tweets posted during natural disaster can be a great source of real time & valuable information for saving and serving the affected people in need. As there is a limited scope of manual interpretation of social media tweets in a situation where immediate action is required. Hence, there is a growing demand of an information retrieval system and related technologies which can extract the meaningful situational and relevant information from these tweets. There exist few issues while extracting the useful information from these tweets.

The enormous number of tweets per unit of time is a is a major issue to handle and pre-process for further analysis and summarization.

Revised Manuscript Received on February 28, 2020.
* Correspondence Author

Dr. Santosh K. Vishwakarma*, is working as Associate Professor in the department of CSE, School of Computing & IT, Manipal University Jaipur, India.

Rishabh Singh Chandel, Master’s Student in Computer Science & Engineering from Gyan Ganga Institute of Technology and Sciences, Jabalpur, India.

Parul Hora, Masters Student in Computer Science & Engineering Department of Gyan Ganga Institute of Technology and Sciences, Jabalpur, India.

Feature Extraction from the tweets is another major issue to deal with this kind of short messages.

During some natural disasters like events, the tweets are of diverse category based on its context; which triggers to the need of tweets classification for extraction of useful and relevant information [5]. It is also challenging to verify the credibility of the information extracted from microblogs from other reliable sources [6].

The above problem has also been listed in one of the tracks of FIRE-2018 (Forum of Information Retrieval & Evaluation). The track known as Information Retrieval from Microblogs during Disasters (IRMIDIS) focused on the identification of factual or fact-checkable tweets and supporting news article for each fact-checkable tweet. The data consists of around 50, 000 microblogs (tweets) from Twitter and 6,0000 news articles, that were posted during the Nepal earthquake in April 2015. One of the tasks was to identify factual or fact-checkable tweets.

These microblogging posts posted by the common people contains much meaningful and important situational information which is usually mixed up with the large amount of not so useful general conversation of the people including their views, thinking, sympathetic messages, wishes to the disaster victims and so on.

This paper addresses the above issues by making use of the research data present in the IRMIDIS track of FIRE-2018. The track which consists of varying microblogs are used for training of model to build a classifier which can automatically recognize need and availability-based tweets. The classifier thus obtained is useful for getting situational information.

The rest of the paper is organized as follow: section-2 presents the literature review related to the present work. Section 3 explains the methodology that we used in collecting, pre-processing, feature extraction, disaster lexicon creation, and classification scheme used for annotating the tweets. Section 4 describes the architecture of the proposed system. Section 5 describes the result followed by the discussions. The major findings and future scope are covered in Section 6 of the paper.

II. RELATED WORKS

In recent years, few researchers & machine learning forums have focused on the issues related to tweets posted during natural disasters and proposed models for feature extraction, categorization, summarization, classification, etc. During the disaster events like flooding, earthquake, fires, the social media including the microblogging sites plays a vital role for providing the situational information [7][8].

The case study of the microblogging effect in 2010 Yushu earthquake in China understand the roles played by microblogging systems in response to major disasters and enabled us to gain insight into how to harness the power of microblogging to facilitate disaster response [9].

Win et al. [5] proposed a system which successfully annotated the Myanmar Earthquake data at 75% accuracy on average. The method combines feature extraction using NLP and machine learning approach to obtain the annotated datasets to improve disaster response efforts. Their method outperforms the standard bag of word model and neural word embedding model.

Verga et al. [10] proposed a method to discover matches between problem reports and aid messages from tweets in large-scale disasters. They concluded that the performance of the problem-aid matching can be improved with the usage of semantic orientation of excitation polarities and trouble expressions.

With context to the situational awareness Verma et al. [11] shows that the low-level linguistic features such as subjectivity, personal/impersonal style, and register substantially performs well at identifying tweets.

The working notes of the FIRE-2016 task [12-14] also summarizes the result obtained after various experiments with feature extraction and principal component analysis. They experimented with the set of previously collected tweets on Nepal Earthquake 2015 and provided answer to the 7 queries that were given in the traditional TREC [15] format.

Many research data reveals that the 85% of the most trending topics over the Twitter are the news headlines. People are actively found searching for breaking news and real-time contents during any crises and natural disaster on Twitter [16].

Another important research done by [17] concludes that how social media platforms are extensively being used in varied geographic regions around the globe, and the most important task in current scenario is that which analytical methods are most suitable to reproduce and transfer methods from one geographic location to another.

III. METHODOLOGY

A) Task Description

The FIRE track consists of microblogs with approx. 50000 posted during Nepal Earthquake of 2015 in the TREC format. The experiments have been performed on two specific type of microblogs or tweets as following;

Need Tweets - Tweets which informs or points to the need or requirement of some specific resource needed by the people in the disaster stuck area like: -Food, Water, Medical aid, Shelter etc. The tweets which do not directly specify the need but point to some scarcity or non-availability of some specified resource are also included in this category.

Availability Tweets: Tweets which informs or points to the availability of some specified resources. These type of tweets class includes microblogs belonging to two different section like one which inform about the potential availability , such as resources being transported or dispatched to the disaster struck area, and also the other class which informs about the actual availability in the disaster- struck area for example items and resources being distributed on the site of disaster etc.

With the main aim of identifying the need and availability tweets for effective coordination of post disaster relief operation, we have devised a pattern matching IR technique which will match the need tweet with the corresponding availability tweet and hence identifying the actionable information discarding the unusual information.

The proposed system operates on the four main components;

Microblog pre-processing: As the microblogs are written by the common people and hence, they include lot of unnecessary information and noise, so in this stage data filters are applied on the tweets to get rid of unwanted data.

Query construction: Queries are constructed to find specific data by filtering specific criteria.

Scoring of tweets: In this a model is applied the new data after the process of query construction known as scoring. The scored data is the actual data in which the model has been applied.

Final filtering: After each of the tweets get assigned by a score, we then applied some heuristic threshold values to get such tweets which will be our main and refined data source which is of good quality needed for such topic.

B) Tweet Pre-processing

The following steps are adopted for the processing of microblogs for further processing;

Punctuation Removal– We have removed the punctuations from each of the tweets, not giving extra importance to the useless symbols and signs.

Case Folding- All the upper-case letters present in the tweets are converted to the similar lower case.

Stop Words Removal-We have created a bag of such stop words which are of least importance and holds no significance and are used only for semantic purposes to frame a sentence. Afterwards the stop words included in the tweets have been filtered and removed, to leave the meaning full data only relevant to our experiment.

Non-ASCII character removal– All the non-ASCII characters including special symbols and emoticons included in the tweets have been removed from the tweets.

Constructing bag of words- Now in this step we have created a bag of words which will be identified by the tweet ids, this will give every tweet a unique identification and will help us in tracking the information in the next step. This is usually done by splitting the above filtered tweet into words which are then grouped into a set of words. Each set will have collection of distinct words that are included in the tweet.

C) Query Construction

Whatever tweets we have either manually made or used, we divided the topic into three fields, namely the title, description and narratives. Each of the tittle contained many keys. Description part contains the detailed one liner information about the user's information needs, whereas the narratives are the paragraph wise long description of the user's needs.

We have assigned each of the topic a unique id which will be used to uniquely identify it in the submission stage. On the whole query construction part consists of the two different phases.

Keyword Extraction: As only the noun part of the tweets only holds the most required information, so we have used filtering which is based on (POS) tagging.

Giving weight to keyword: All the topics present can be broadly categorized into two groups as per our requirement they are need tweets (requirement) or availability tweets. These two have been assigned more weightage.

D) Scoring:

After query construction we have used 2 different scoring methods as SVM and weight modification with information gain applied with SVM.

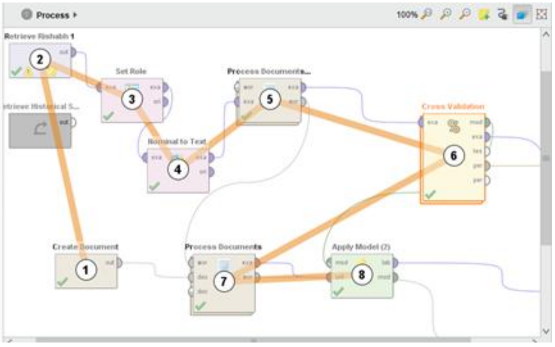


Figure 1: Design workflow SVM Model without term weight change

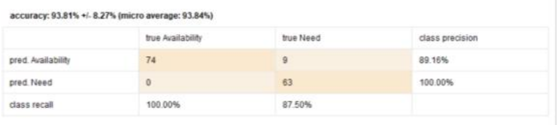
E)Final Filtering:

After the scoring of the tweets according to the relevance to our each of the topic, we have opted a heuristically set threshold-based filtering method to choose most relevant tweets. The threshold has been set in our case to 0.50. which means that the tweets which are having more score than this heuristically set threshold 0.50 are useful and considered and relevant and they are used and submitted. The tweets which are below this threshold has been discarded and termed as not useful for our work.

IV. RESULT AND ANALYSIS

The model has been implemented with RapidMiner, a well-known platform for data mining and machine learning experiments. It allows to design various data analysis processes with the use of operators. Additional functionality can be added to the RapidMiner process.

As discussed in the scoring section of Methodology, the first process consists of designing the model by use of support vector machine. The results as obtained are shown in Figure 2 below.

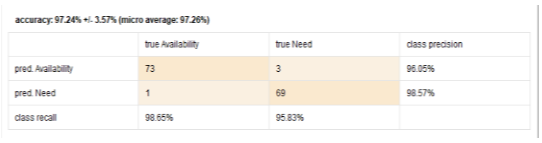


accuracy: 93.81% +/- 8.27% (micro average: 93.84%)

	true Availability	true Need	class precision
pred. Availability	74	9	89.16%
pred. Need	0	63	100.00%
class recall	100.00%	87.50%	

Figure 2: SVM Model with no term weight change

The above model shows approx. 89% of accuracy for the Availability class and 100% accuracy for the Need class. The performance measure used in the experiments is accuracy, which is a measurement of how correct the predictions are made by the system. When the same dataset is applied after the weight modification of the terms using information gain method, the accuracy of the model is increased to 97%, which is a marginal growth compared with the no term weight change. The above results are illustrated in Figure 3 below.



accuracy: 97.24% +/- 3.57% (micro average: 97.26%)

	true Availability	true Need	class precision
pred. Availability	73	3	96.05%
pred. Need	1	69	98.57%
class recall	98.65%	95.83%	

Figure 3: SVM Model with term weight change using Information Gain

The proposed approach for modelling of microblogs has been compared using the SVM classifier as a base model with no change in the term weight and change in the term weight using the information gain term modification method. Table 1 gives the comparative results with all these measures and further it has been plotted in Figure 4.

Table 1: Comparison of the Result

Evaluation Parameters	SVM without term weight modification	SVM with term weight modification using IG
Accuracy	93.81	97.24
Classification Error	6.19	2.76
Class Precision	94.58	97.31
Class Recall	93.75	97.24

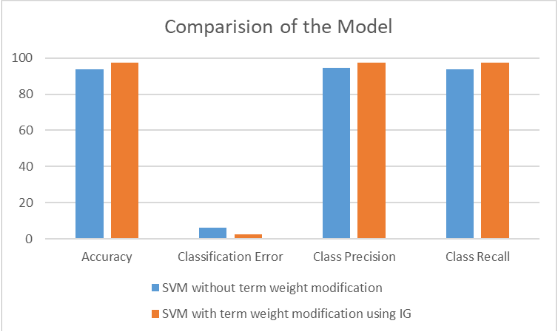


Figure 4: Comparison of the model

The above result shows that applying some term weight modification scheme such as information gain plays a vital role in increasing the efficiency of the model. Various base methods has been used in the proposed approach but the promising results has been obtained when the SVM classification scheme is applied with the microblog database.

V. CONCLUSION AND FUTURE SCOPE

The work provided a brief discussion on the approach to FIRE-2015 microblogs track. We have observed that the traditional method of information retrieval from microblogs tweets can be enhanced much efficiently by including some more and better filtering and scoring algorithms in addition or place of the previously done work. We have completed our objective with much precise score of both identification of the need tweets and availability tweets and then matching the need tweets with the availability tweets.

REFERENCES

1. Antoniou, Natassa, and Mario Ciaramicoli. "Social media in the disaster cycle useful tools or mass distraction?." In *International Astronautical Congress*. 2013.
2. Mathbor, Golam M. "Enhancement of community preparedness for natural disasters: The role of social work in building social capital for sustainable disaster relief and management." *International Social Work* 50, no. 3 (2007): 357-369.
3. Moumtzidou, Anastasia, Stelios Andreadis, Ilias Gialampoukidis, Anastasios Karakostas, Stefanos Vrochidis, and Ioannis Kompatsiaris. "Flood relevance estimation from visual and textual content in social media streams." In Companion Proceedings of the The Web Conference 2018, pp. 1621-1627. International World Wide Web Conferences Steering Committee, 2018.
4. Murthy, Dhiraj. Twitter. Cambridge, UK: Polity Press, 2018.
5. Win, Si Si Mar, and Than Nwe Aung. "Target oriented tweets monitoring system during natural disasters." In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pp. 143-148. IEEE, 2017.
6. Basu, Moumita, Saptarshi Ghosh, and Kripabandhu Ghosh. "Overview of the FIRE 2018 track: Information Retrieval from Microblogs during Disasters (IRMiDis)." In *Proceedings of the 10th annual meeting of the Forum for Information Retrieval Evaluation*, pp. 1-5. ACM, 2018.
7. Cameron, Mark A., Robert Power, Bella Robinson, and Jie Yin. "Emergency situation awareness from twitter for crisis management." In Proceedings of the 21st International Conference on World Wide Web, pp. 695-698. ACM, 2012.
8. Neubig, Graham, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. "Safety Information Mining—What can NLP do in a disaster—." In Proceedings of 5th International Joint Conference on Natural Language Processing, pp. 965-973. 2011
9. Qu, Yan, Chen Huang, Pengyi Zhang, and Jun Zhang. "Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake." In Proceedings of the ACM 2011 conference on Computer supported cooperative work, pp. 25-34. ACM, 2011.
10. Varga, István, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. "Aid is out there: Looking for help from tweets during a large scale disaster." In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1619-1629. 2013.
11. Verma, Sudha, Sarah Vieweg, William J. Corvey, Leysia Palen, James H. Martin, Martha Palmer, Aaron Schram, and Kenneth M. Anderson. "Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency." In Fifth International AAAI Conference on Weblogs and Social Media. 2011.
12. Trishnendu Ghorai. An information Retrieval System for FIRE 2016 Microblog Track. In working notes of FIRE 2016- Forum for Information Retrieval Evaluation.
13. Roshni Chakraborty and Maitry Bhavsar : Information Retrieval from Microblogs during Disasters FIRE 2016 Microblog Track. In working notes of FIRE 2016- Forum for Information Retrieval Evaluation.
14. Saptarshi Ghosh and Kripabandhu Ghosh. Overview of the FIRE 2016 Microblog track: Information Extraction from Microblogs Posted during Disasters. In Working notes of FIRE 2016- Forum for Information Ritrieval Evaluation, Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings. CEUR-WS.org, 2016.
15. https://trec.nist.gov/
16. Zahra, Kiran, Muhammad Imran, Frank O. Ostermann, Kees Boersma, and Brian Tomaszewski. "Understanding eyewitness reports on Twitter during disasters." (2018): 687-695.
17. Zahra, K., Ostermann, F. O., and Purves, R. S. (2017). "Geographic variability of Twitter usage characteristics during disaster events". In: Geo-spatial information science 20.3, pp. 231–240.

AUTHORS PROFILE

Dr. Santosh K. Vishwakarma, is working as Associate Professor in the department of CSE, School of Computing & IT, Manipal University Jaipur. He completed his bachelor's and master's degree in Computer Science & Engineering. He is a doctorate in the field of Information Retrieval. His teaching specialization includes database management system, operating system, compiler design and theory of computation. He holds 15 years of Teaching Experience in reputed Institute. His research interest includes data mining, text mining, predictive analysis. He has been invited in various national and international forums for delivering sessions on databases, big data, predictive analysis and machine learning algorithms.

Rishabh Singh Chandel, He is a Master's Student in Computer Science & Engineering from Gyan Ganga Institute of Technology and Sciences (Jabalpur) affiliated to Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal. He has graduated with a Bachelor's of Engineering degree in Computer Science Stream, in 2015 from Gyan Ganga College of Technology (Jabalpur). He is currently operating his own institute providing education and imparting knowledge to the students in Surat (Gujarat). He has a keen interest in new technology and Data Mining field.

Parul Hora, She is a Masters Student in Computer Science & Engineering Department of Gyan Ganga Institute of Technology and Sciences, Jabalpur affiliated to Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal. She graduated with a Bachelor's of Engineering in Computer Science Department, in 2015. She is also a teacher assistant in St. Xavier's Group of Schools, Jabalpur. She has keen interest in Computer and Data Mining.