Disaster Management using Ontology Feature

Parul Hora
GGITS, Jabalpur

Neha Sheth
Manipal University Jaipur

Santosh K. Vishwakarma
Manipal University Jaipur

ABSTRACT

The digital transformation has witnessed an exponential growth in the recent years. This transformation has touched every instance of human life. The current generation rigorously rely in the platform of information storage & retrieval. This originates an ample opportunity for designing and developing systems for societal benefits. The importance of social networking forums has a vital role in our life. The usage of the above websites is frequent towards maintaining identity, keeping connect with the friends, updating personal & professional information, etc. They have increasingly infused itself into daily life. In recent years, one of the important areas of research is oriented towards from the social networking websites in different categories. This paper represents the work done with an open research dataset known as Microblog Track provided by Forum of Information Retrieval & Evaluation (FIRE). The task provided by the FORUM is to develop a suitable model for the identification of tweets. The training dataset consists of two predefined labels, known as need and availability. In this paper, the prediction rate has been optimized by using the term weighting models before applying the classifiers. 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, Microblogging sites, Twitter, Precision, Recall

1. INTRODUCTION

According to the UN, a disaster is a serious disruption to the functioning of a community, which causes human, material, economic and environmental losses beyond a community's ability to cope. [1] In such abrupt situations, it becomes imperative for individuals, communities, and governments to refurbish their disaster management technologies, especially information systems that enable communication of dynamic data from the site of the event as well as predictive technologies that can rapidly process new data and generate insights for prompt decision making.

Technologies like unmanned aerial vehicles or drones that capture the damage in real time, remote sensing satellites help detect anomalies in weather and seismic parameters (e.g., LandsatTM for vegetation mapping, Meteosat for rainfall mapping and AMSR-E for flood mapping and forecasting. [2]) and data visualization platforms have played an important role in disaster management. In the digital era of social media, communication and collaboration has reached new heights and during times of crisis, people often turn to social media and micro-blogging platforms to share their needs, concerns, and resources with each other [3]. As a result, huge amounts of heterogeneous data are constantly being collected which can substantiate big data analysis and predictive modeling to better prepare for disasters [4]. A big challenge in this area is the organization of this unstructured data as well as subsequent analysis and modeling.

The disaster management cycle includes mitigation, preparedness, response and recovery. All four components of the cycle can benefit from AI systems that utilize natural language processing and semantic technologies. Because of the diversity of structured and unstructured data collected, semantic methodologies can be used to extract meaning from data, encode and decode expressions, perform classification and generate knowledge graphs. Furthermore, natural language processing and text classification technologies allow rapid collaboration between communities, such as providing leads for food supply, distribution of resources and coordination between people in need.

Other challenges encountered by new technologies is the unavailability of real-time data due to the uncertainty of disasters. Setting up long-term technological strategies to prepare for disasters and quick mitigation procedures are crucial for prompt response during times of crisis. For example, analysis of existing semantic data to build accurate systems that allow for quick action. It is also important to generate dynamic information for people in need, such as real-time streamlining of posts and micro-blogs by people and communities so that the necessary aid can be provided well in time.

The idea behind ontology is well applied in this approach as sharable and reusable knowledge are used for pattern semantics and predicting with the models.

2. REVIEW OF LITERATURE

The research on the information retrieval field has recently focused on many new areas that makes direct impact to the human life. The key information extracted and analyzed in the case of natural disaster have witness a huge attention of the researchers. The locality of reference model in terms of spatial information plays a key role in the field as mentioned [7][9]. The paper [9] discussed the case study of the natural disaster in China in the form of earthquake. They have given emphasis on how to utilize the microblogging feature to manage better disaster response. The research carried out of win et al. [5] focus on the microblog tweets posted during the Myanmar Earthquake. They applied the feature extraction method with Machine Learning approach for identification of the message types. The bag of word model has also been evaluated for the identification and term weight application by [10]. They reported the problems and aid messages from large scale disasters. The research work also focuses on the semantic orientation of interconnecting the messages. The paper [11] shows that subjectivity, style, register substantially creates valuable impact during the identification of the tweets.

3. METHODOLOGY

The experiments have been carried out with the open research dataset of Forum of Information Retrieval and Evaluation [16]. The track consists of tweets posted during the disaster by

the people who suffered in this incident. There exist more than seventy thousand tweets posted and filtered for the research process. Our experiment workflow is shown in Figure 1.

Figure 1: Workflow of the experiments

The above figure 1 shows the workflow of the experiments. The major steps include tokenization, stop word removal, stemming and generate n-grams during pre-processing of data. Tokenization refers to process of taking character sequence from defined document unit and breaking it into words, symbols, phrases, and numbers called tokens. Stop word removal filter out the words that have no values for retrieval purpose. Stemming perform replacement of all the variations of the words with its root word. The variant words may be plurals, gerund forms, prefixes, suffixes etc. a stem word can represent all its variants that reduces the size of dictionary containing all words of document collection. In our analysis Porter algorithm found best because it produces maximum number of tokens.

After the preprocessing steps, various traditional classifiers are used for training the model. The evaluation metric is analyzed, and the results obtained are discussed in the next section. After the first evaluation with various classifiers as stated above, the term weight has been updated with various extended version of tf-idf model. The changes in the term weight significantly improved the performance of the classifiers specially in the case of Information Gain approach. This method modifies the term weight by using the concept of entropy. The entropy can be shown as

$$\text{Entropy (P)} = -[p1\log(p1) + p2\log(p2) + \ldots\ldots\ldots + pn\log(pn)].$$

Where P = (p1, p2,…., pn) are the probability distribution as the conveyed information through this distribution is called the entropy of P. For example, for 'n' equally probable messages, the probability 'p' is 1/n of each. Hence the message conveys the information as –log2 (p) = log2 (n). It should be noted that for more or greater information, the more uniform should be the probability distribution.

From decision tree context, if the attribute of interest is divided into two classes C1 and C2, then the result can be seen as a message being generated. Entropy provides the measure of message information to be of class C1 or C2. If the record set T is divided into set of 'n' classes such as C1, C2, …….., Cn based on the class attribute, then the information required to determine the class of the item T is

$$\text{Info (T)} = \text{Entropy (P)}$$

Where P= probability distribution, C1, C2, …. , Cn is the partition.

The information gain is used as attribute selection measure by ID3 algorithm of decision tree. For splitting a specific node, the attribute with the highest information gain is chosen.

4. RESULT AND ANALYSIS

For illustrating the data mining and language processing techniques in case of disaster management, we have used Rapid Miner, a data science platform with visual workflow design and automation of model development. It has shown useful results on the sample data used, and the model can be functional for 60 to 600 rows of unlabelled data as well.

As discussed in Methodology, seven classification algorithms are implemented using two main methods wherein term frequency weight has been employed for better results. Cross validation is performed with 10 numbers of subsets or folds (9 for training and 1 for testing). Furthermore, the model is also tested on unlabelled testing data.

We find that without TF-IDF, the most accurate algorithm is KNN (where K=5) with an accuracy of 83% and class recall of 83.33% respectively. The second most accurate algorithm is Neural Network with 200 training cycles and a learning rate of 0.01. It shows an accuracy of 79.67% without TF-IDF weight modification. Respective confusion matrices are built for all the algorithms implemented and several performance measures are used to evaluate the model, namely accuracy, classification error, class precision and class recall.

accuracy: 83.00% +/- 15.75% (micro average: 83.05%)

	true Availability	true Need	class precision
pred. Availability	25	5	83.33%
pred. Need	5	24	82.76%
class recall	83.33%	82.76%	

Figure 2: Accuracy Measure with the FIRE Tweets

Table 1:Classifiers Performance with TF-IDF Model

Algorithm	Accuracy	Classification Error
K-NN	**83.00**	**17.00**
Neural Network	79.67	20.33
SVM	78.00	22.00
Naïve Bayes	66.33	33.67

After the first evaluation with various classifiers as stated above, the term weight has been updated with various extended version of tf-idf model. The changes in the term weight significantly improved the performance of the classifiers specially in the case of Information Gain approach.

Table 2: Applying Term Weight as Information Gain

Algorithm	**Accuracy**	**Classification Error**	**Term weight scheme**
K-NN	86%	14%	IG
Neural Network	81%	19%	IG
SVM	**91%**	**9%**	**IG**
Naïve Bayes	79%	21%	IG

The result as shown in Table 1 and Table 2 differentiate the classifiers accuracy, classification error with the TF-IDF Scheme and with Information Gain scheme. The calculation of entropy and filtering the words increases accuracy with the IG Scheme. We have not shown the results pertaining to other classifiers because their performance was not promising with this scheme.

5. CONCLUSION

The work carried out in the thesis is the approaches towards classifying the tweets from the real-life platform of FIRE-2015 microblogs track. The experiments carried out with the traditional classifiers such as Naïve Bayes and Support Vector Machine. The SVM model is also used with the term weight change for the optimized results. In particular, the main claim behind this thesis work has been how data mining techniques can be utilized to enhance the effectiveness of predicting the survivor based on the tweets posted during the disaster.

The main objective of our work was to discover and design the classifiers that can be applied during the natural disasters such as earthquakes, huge rain falls, floods, tsunamis, etc. The models designed for the experimentation purpose are standalone models. We intend to apply the multimodal approach to find interesting patterns and optimize results. Although, the results obtain in our experiments are very useful for tracing the person or survivor during the natural disasters. We also found that most of the traditional model were not fitted for the classification of tweets and gives low accuracy below 45%, so we discarded such models and included those models who provides satisfactory performance in terms of accuracy and minimizes the classification error.

One of the major findings in our work is that preprocessing of the tweets during natural disasters is one of the essential tasks before the training of the model. The reason behind these is the unwanted number and sequence of characters typed in such situations. The various preprocessing steps performed in our experiments were useful to remove this type of errors and find the useful keywords from the dataset.

6. 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, IliasGialampoukidis, Anastasios Karakostas, Stefanos Vrochidis, and IoannisKompatsiaris. "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 ThanNwe 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, YuichirohMatsubayashi, 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, KentaroTorisawa, Chikara Hashimoto, KiyonoriOhtake, 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, LeysiaPalen, 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] TrishnenduGhorai. An information Retrieval System for FIRE 2016 Microblog Track. In working notes of FIRE 2016- Forum for Information Retrieval Evaluation.

[13] Roshni Chakraborty and MaitryBhavsar : 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] http://fire.irsi.res.in/fire