Automatic Identification and Ranking of Emergency Aids in Social Media Macro Community

Bhaskar Gautam
Data Informatics and Analytics Lab
National Institute of Technology, Karnataka
P.O. Srinivasnagar
Surathkal-575025, India 575025
bhaskar.gautam2494@gmail.com

Annappa Basava
Data Informatics and Analytics Lab
National Institute of Technology, Karnataka
P.O. Srinivasnagar
Surathkal-575025, India 575025
annappa@ieee.org

ABSTRACT

Online social microblogging platforms including Twitter are increasingly used for aiding relief operations during disaster events. During most of the calamities that can be natural disasters or even armed attacks, non-governmental organizations look for critical information about resources to support effected people. Despite the recent advancement of natural language processing with deep neural networks, retrieval and ranking of short text becomes a challenging task because a lot of conversational and sympathy content merged with the critical information. In this paper, we address the problem of categorical information retrieval and ranking of most relevance information while considering the presence of short-text and multilingual languages that arise during such events. Our proposed model is based on the formation of embedding vector with the help of textual and statistical preprocessing, and finally, entire training 2,100,000 vectors were normalized using feed-forward neural network for need and availability tweets. Another important contribution of this paper lies in novel weighted Ranking Key algorithm based on top five general terms to rank the classified tweets in most relevance with classification. Lastly, we test our model on Nepal Earthquake dataset (contains short text and multilingual language tweets) and achieved 6.81% of mean average precision on 5,250,000 unlabeled embedding vectors of disaster relief tweets.

CCS CONCEPTS

•**Information systems** →**Data stream mining;** *Data Cleaning;* Clustering and Classification; •**Human-centered computing** →Social tagging systems;

KEYWORDS

Automatic Information Retrieval Model, Transliteration, Multilingual Language, Embeddings, Multi-layer Perceptron, w-Ranking Key Algorithm

1 INTRODUCTION

The branch of Information-driven and computational social science research is growing in importance [9], and with this trend, there is a need of common fully-automatic information retrieval model to facilitate a robust and reproducible research environment [6]. Initially, Online Social Networks (OSNs) aims to allow users to communicate, connect with others and share content [3] but nowadays the OSNs allows its users to broadcast personal thoughts and

content. During natural or man-made disasters, these social networks provide the rich source of near-real-time data streams which helps in the formation of emergency situation summarization [15]. However, increasingly usage of these social media tweets for aiding relief operations during disaster events such critical information is usually submerged with a lot of conversational content, such as sympathy for the victims of the disaster. Hence, automated Information Retrieval techniques are needed to extract availability and absence of any kind of resource to save life.

In this paper, we focus on automatic classification and ranking problem [12, 16] for two types of critical information, named as *Need-tweets* and *Availability-tweets*. Our goal is to improve classification methods and return the tweet in most relevance order with the class while considering the fact of the presence of short-text and multilingual language in OSNs. A glimpse of the challenge described in FIRE 2017 Information Retrieval from Microblogs during Disasters Track [1] and also discussed in [1]. Training and evaluation of the entire model is done using 70,000 tweets which were posted during the 2015 Nepal earthquake. However, including the need and availability tweets corpus also include some redundancy such as the presence of similar tweets and all such false tweets were removed by Track organizers as discussed in [8] in order to preserve uniqueness of tweet labels. The purpose of this paper is to consider the highlighted problem, and the trade-offs that they entail. Our paper makes the following contributions (1) We describe the methodology and (2) release the pseudo code in order to replicate the results of entire paper. The remainder of this paper is structured as follows: in Section 2, we present the Feature resolution methods used to create a feature vector for neural network model. Section 3 presents methodology followed in the formation of entire information retrieval model. Section 4 discusses the task required to form classification model. Section 5 discusses the weighted ranking key algorithm and Section 6 depicts about the evaluation and result of the model while comparing with another team. Finally the paper concludes in Section 7 with discussing some of the conceivable use cases.

2 FEATURE RESOLUTION METHODS

In this section, we discuss the Feature Resolution Methods which helps in the formation of embedding vectors proposed to consumed by our classification model. The methods are - **Textual preprocessing** and **Statistical Preprocessing** which exploits the public

---

[1]https://sites.google.com/site/irmidisfire2017/

information of social network users with the help of Stanford NLP Toolkit [2] while preserving privacy. A detailed description of each method is elaborated in following sections.

2.1 Textual Preprocessing

The goal of this phase is to make all tweets in a particular semantics some of them are discussed in [7, 14] considering the tweet which are written in a language (specifically English, Hindi, Nepal, and language which is the combination of Hindi + English). Tweets are represented as vectors of features, each feature being a word unigram, by performing standard text preprocessing operations whose highlights are depicted as follows:

- Removing non-ASCII characters.
- Removing all the stopwords (also for Hindi and Nepali).
- Normalizing all the text of tweets into lowercase.
- ASCII transliterations of Unicode text (Tweet).
- Separation of sentence into tokens (words)
- Stemming of social media tweets using the standard Porter stemmer.

2.2 Statistical Preprocessing

During Disaster-related crises, the social media tweets which belonged to the same context generally submerged with interrelated keywords which summarized the entire tweet (are discussed in Table 1) and referred in the rest of the paper. These keywords consist of tweets which describe the resource availability status and moreover the need of the resources.
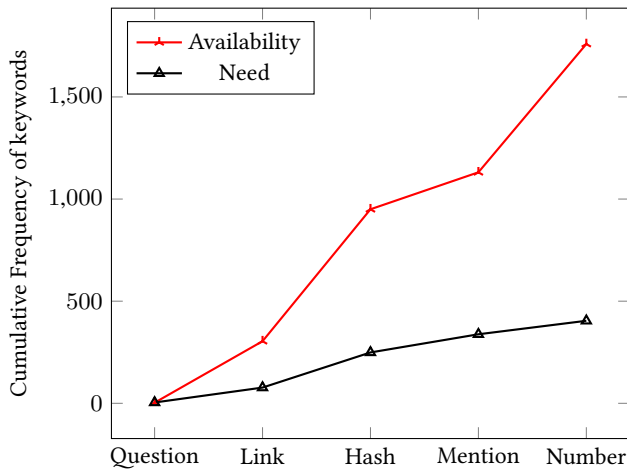

Figure 1: Statistical Relationship of top 5 Features among different social media macro community category in entire training corpus

The contribution of these keywords are described in Figure 1, where *Availability Tweets* have the maximum contribution. To have reflection of these tokens in tweet, we generalize some tokens by replacing ? by the token QUESTION, hyperlinks by the token LINK, Hashtag by the token HASH, MENTION token is the replacement

---

of (@username) and lastly, any number of the tweets are replaced by NUMBER.

3 METHODOLOGY

Most of the users in Online Social Network (specifically, Twitter) share personal information, random thoughts, opinions/complaints and facts [11]. Moreover, with in near-real-time these social network explodes with an activity whenever any calamity takes place [2]. Considering above statements as a hypothesis for this paper, we combine the discussed feature resolution methods to create a fully-automated system, named as Categorical Information Retrieval (IR) and outlined in architecture Diagram 2. Given the social networking post (tweet, that are associated with the related hashtag), the Categorical IR initiates with the formation of two feature vectors termed as *Textual preprocessing and Statistical Preprocessing*.
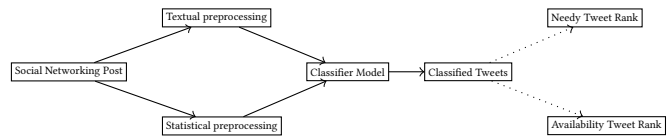

Figure 2: Architecture and Methodology of Categorical IR

The formation of these feature vectors are essential as it normalizes the semantics of the given post. These preprocessed posts are transformed into embedding [10] vectors of size 100 and 5 in length respectively. Once the initialization of entire embedding are done, our classification model (further elaborated in Section 4) precisely able to categorize any tweets with respect to its relevance with the help of these embeddings. After enforcing these requirements, our Ranking algorithm will precisely rank the classified tweet irrespective of any language within a relevance class.

4 CLASSIFICATION MODEL

Classification model initiates with the formation of words embeddings for every social media tweet using Textual preprocessing with the help of Doc2vec model[10], well suited for variable length and short length text which outperforms the Word2vec model and probabilistic latent semantic analysis model approach discussed in [5, 8, 13]. Now, each of the word embedding (of size 100 in length) combines another feature vector which was formed using the frequency of each token (discussed in Section 2.2) available in a tweet. After preprocessing, we have one entire embedding vector of length 105 which will be the final representative of the respective class.

Social media tweets in entire training corpus are transformed into 21,00,000 vectors (each of length 105) in which the absence of most similar tweets are ensured by the track organizers. These embeddings will be the input for multilayer preceptron (a feedforward Artificial Neural Network model). Multilayer preceptron model runs upto 300 iterations with an "ReLu" as a activation function. During the training of model, entire training embedding vectors (2,100,000) were cross-validated with K-folds (k=10), and gives

---

Table 1: Description of Top five keywords available in Corpus

Keyword	Description	Annotation
Question	Request for any resources	β_{need}
Link	Offer availability of resource using portal	$\gamma_{need}, \beta_{availability}$
Hash	Broadcast availability of resource using all popular Hashtags	$\gamma_{availability}$
Mention	Offer some help while publicizing particular community (Political Party etc.)	$\alpha_{availability}$
Number	With Mobile number offer help (Ambulance etc.)	α_{need}

maximum precision and recall in comparison with Naive Bayes, Logistic Regression, SVM even with Random Forest including F-Score.

5 CATEGORY BASED RANKING

Classification model effectively classifies each social media tweet to the respective class and each classified tweet along with the prediction score (termed as $score_{need}$ and $score_{availability}$) are collected separately as depicted in Figure 2. With the weighted(w)-Ranking Key algorithm, we transform the probability of tweet belong to relevance class with the rank score (termed as $Rank_{score}$). Ranking of tweets based on $Rank_{score}$ helps in prioritizing tweets with most relevant to class considering the presence non relevance tweets.

5.1 w-Ranking Key Algorithm

w-Ranking Key Algorithm proceeds with the initialization of two buckets of top 20 keywords from each relevance class of tweets available in "Classified Tweet" using Twitter LDA (TLDA) topic modeling model which already outperforms the state-of-the-art model on news corpora discussed in [4, 17] and in rest of the paper, we refer these bucket as a $K_{20} - Need$ and $K_{20} - Availability$ respectively. To provide more weights to tweets which contain keywords as described in Table 1, we have created two more small buckets labeled them as $H_5 - Need$ and $H_5 - Availability$ respectively. Moreover, each keyword is classified into most relevance class and annotation of each keyword with respect of class also shown in Table 1. You may note that each keyword can be assigned to more than one category, as the corresponding keyword could belong to more than one category.

For each classified tweet, algorithm makes count the presence of each annotated keyword with α_1, α_2, and α_3. We pool the relevance class score with the updated relevance class score. Now, if the similarity score between the word available in each tweet and the tokens available in buckets (H_5 and K_{20} of category) exceeds a certain high confidence threshold C, we exponentially upgrades the pool score with itself (and hence it provides the more weight to keyword which is most relevant). For our similarity metric between keywords, Levenshtein Distance (named after Vladimir Levenshtein, 1965) is preferable among another more obvious candidate in string similarity metrics considering the presence of short text language used in a tweet. Moreover, the algorithm assigns negative weights to the tweet which contains most common keywords available in a most non-relevance tweet such as "quake" and

any floating-point number. Entire pseudo code of algorithm is discussed in Appendix A.

6 EVALUATION AND RESULTS

Primarily, the motivation for this paper is to get a quick overview of the current fully-automated state-of-the-art models for disaster reliefs for various countries. We reviewed several Information Retrieval (IR) models but none of them have the better precision for short text and multilingual languages. In this section, we highlight the precision and recall of our proposed fully-automatic IR model for disaster reliefs (obtained with the prediction of class and ranking for 50,000 tweets). The performance evaluation of model
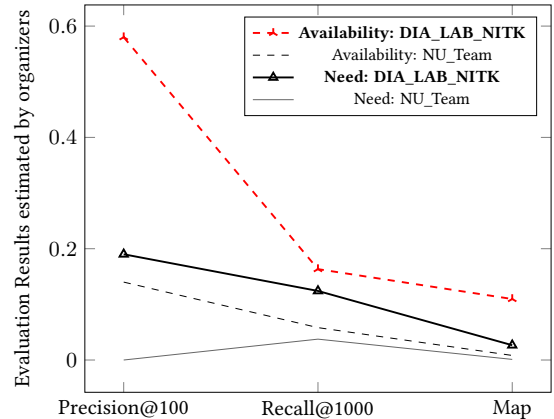

Figure 3: Performance evaluation of classification model (including Ranking Key Algorithm) in terms of Precision with top 100 ranked tweets and Recall with top 1000 ranked tweets including comparison of results with another team.

shown in Figure 3 while comparing the evaluation results with another Team (NU_Team) participated in track. Different evaluation metrics (such as precision, recall) and measures were used on top 100 and 1000 tweets (ranked by our algorithm) respectively. Finally with the help of mean average precision results were evaluated while comparing with a gold-standard dataset of Nepal Earthquake.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we addressed the problem of categorical information retrieval and ranking model for disaster relief while considering the presence of short-text and multilingual languages. In such a

scenario, we can expect to have a lot of conversational and sympathy content submerged with the critical information. Moreover, most of the critical information on Online Social Media regarding disaster relief are either in form of short-text or in local regional language. Hence, Our fully-automated *Categorical information retrieval model for disaster relief* integrates textual and statistical pre-processing methods. Through the combination of textual and statistical embeddings we train our model with a multi layer preceptron model with the presence of 76% Availability Tweets and 23% of Need Tweets achieved 6.81% of mean average precision on 50,000 unlabeled disaster relief tweets. Categorical model can be used in evaluation of number available resources with need of resource ultimately it helps in reducing the load on public emergency services and coordinate social effort to save lives. Our experiments on Twitter data from Nepal Earthquake dataset strongly indicates that the proposed fully-automated supervised approach could most often result in improving the accuracy as compared to state-of-the-art approaches and produces the similar result on other dataset even on real time streaming. We plan to extend our approach with deep residual networks and with this strategy it is expected to result in further improvement in precision and recall of architecture.

Categorical model can be used in evaluation even mapping of number available resources with need of resource and with such mapping it reduces the load on public emergency services and coordinate social effort to save lives. Our experiments on Twitter data from Nepal Earthquake dataset strongly indicates that the proposed fully-automated supervised approach could most often result in improving the accuracy as compared to state-of-the-art approaches and produces the similar result on other dataset even on real time streaming. We plan to extend our approach with deep residual networks and with this strategy is expected to result in further improvement in precision and recall of architecture.

A   APPENDIX

REFERENCES

ALGORITHM 1: Ranking Key Overview

$\lambda_{Need} \leftarrow \max(score_{need})$
$\lambda_{availability} \leftarrow \max(score_{availability})$
Function *generate* *(Given Category, Generate Rank$_s$core)*
    for *each tweet $\in$ tweets$_{DataBase}$* **do**
        if *tweet$_{category}$ $\in$ category* **then**
            $\alpha_1 \leftarrow$ Number of $\alpha_{category} \in tweet$;
            $\alpha_2 \leftarrow$ Number of $\beta_{category} \in tweet$;
            $\alpha_3 \leftarrow$ Number of $\gamma_{category} \in tweet$;
            for *each word $\in \left(K_{20-category} \text{ and } H_{5-category}\right)$*
            do
                $score_{category} \leftarrow score_{category} * \lambda_{category}$;
                if *similarity (\{FP or "quake"\}, tweet) > C* **then**
                    $score_{category} \leftarrow score_{category} * -1$;
                end
            end
            $Rank_{score} \leftarrow score_{category}$;
        end
        return $\left(\alpha_1 + \alpha_2 + \alpha_3\right) * Rank_{score}$;
    end
end

---

[7] Ali Hürriyetoglu and Nelleke Oostdijk. 2017. Extracting Humanitarian Information from Tweets.. In *SMERP@ ECIR*. 73–84.

[8] Prannay Khosla, Moumita Basu, Kripabandhu Ghosh, and Saptarshi Ghosh. 2017. Microblog Retrieval for Post-Disaster Relief: Applying and Comparing Neural IR Models. *CoRR* abs/1707.06112 (2017). http://arxiv.org/abs/1707.06112

[9] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational Social Science. *Science* 323, 5915 (2009), 721–723. https://doi.org/10.1126/science.1167742 arXiv:http://science.sciencemag.org/content/323/5915/721.full.pdf

[10] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. *CoRR* abs/1405.4053 (2014). http://arxiv.org/abs/1405.4053

[11] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. 2010. Is It Really About Me?: Message Content in Social Awareness Streams. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. ACM, New York, NY, USA, 189–192. https://doi.org/10.1145/1718918.1718953

[12] Tahora H. Nazer, Fred Morstatter, Harsh Dani, Huan Liu, undefined, undefined, undefined, and undefined. 2016. Finding requests in social media for disaster relief. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* 00 (2016), 1410–1413. https://doi.org/doi.ieeecomputersociety.org/10.1109/ASONAM.2016.7752432

[13] Dat Tien Nguyen, Shafiq R. Joty, Muhammad Imran, Hassan Sajjad, and Prasenjit Mitra. 2016. Applications of Online Deep Learning for Crisis Response Using Social Media Information. *CoRR* abs/1610.01030 (2016). arXiv:1610.01030 http://arxiv.org/abs/1610.01030

[14] Hemant Purohit, Carlos Castillo, Fernando Diaz, Amit Sheth, and Patrick Meier. 2013. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday* 19, 1 (2013). https://doi.org/10.5210/fm.v19i1.4848

[15] Koustav Rudra, Siddhartha Banerjee, Niloy Ganguly, Pawan Goyal, Muhammad Imran, and Prasenjit Mitra. 2016. Summarizing Situational and Topical Information During Crises. *CoRR* abs/1610.01561 (2016). arXiv:1610.01561 http://arxiv.org/abs/1610.01561

[16] Koustav Rudra, Subham Ghosh, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh. 2015. Extracting Situational Information from Microblogs During Disaster Events: A Classification-Summarization Approach. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. ACM, New York, NY, USA, 583–592. https://doi.org/10.1145/2806416.2806485

[17] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. *Comparing Twitter and Traditional Media Using Topic Models*. Springer Berlin Heidelberg, Berlin, Heidelberg, 338–349. https://doi.org/10.1007/978-3-642-20161-5_34

[1] Moumita Basu, Saptarshi Ghosh, Kripabandhu Ghosh, and Monojit 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] Prateek Dewan, Varun Bharadhwaj, Aditi Mithal, Anshuman Suri, and Ponnurangam Kumaraguru. 2016. Visual Themes and Sentiment on Social Networks To Aid First Responders During Crisis Events. *CoRR* abs/1610.07772 (2016). http://arxiv.org/abs/1610.07772

[3] Mats Eriksson and Eva-Karin Olsson. 2016. Facebook and Twitter in Crisis Communication: A Comparative Study of Crisis Communication Professionals and Citizens. *Journal of Contingencies and Crisis Management* 24, 4 (2016), 198–208. https://doi.org/10.1111/1468-5973.12116

[4] Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016. Using Word Embedding to Evaluate the Coherence of Topics from Twitter Data. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, New York, NY, USA, 1057–1060. https://doi.org/10.1145/2911451.2914729

[5] N. Gupta, Abhinav K. R., and Annappa. 2013. Fuzzy sentiment analysis on microblogs for movie revenue prediction. In *2013 International Conference on Emerging Trends in Communication, Control, Signal Processing and Computing Applications (C2SPCA)*. 1–4. https://doi.org/10.1109/C2SPCA.2013.6749350

[6] Maria Han Veiga and Carsten Eickhoff. 2016. A Cross-Platform Collection of Social Network Profiles. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, New York, NY, USA, 665–668. https://doi.org/10.1145/2911451.2914666