Section A-Research paper

Word Embedding based Proposed Algorithm for Information extraction from microblogs for emergency relief and preparedness

Harshadkumar Prajapati[1], Hitesh Raval[2]

1 Research Scholar, Faculty of Computer Science, Sankalchand Patel college of Engineering, Sankalchand Patel University, SK Campus, Visnagar, Gujarat.

2 C. J. Patel college of computer studies, Sankalchand Patel University, SK Campus, Visnagar, Gujarat.

Abstract : Websites for microblogging are crucial sources of current information during any disasters, whether they are man-made or natural. Designing and testing Information Retrieval (IR) systems or algorithms that pull information from microblogs is so crucial. It is also necessary to retrieve the information in accurate and fast manner. In this research paper one word embedding based algorithm is proposed. The algorithm is experimented with more than 50000 tweets related to natural disaster of Nepal earth quake. The algorithm is compared with well known machine learning based algorithms such as Decision Tree, Random Forest and Naïve Bayes. The result of proposed algorithm and all other well-known algorithms is measured in precision, recall, F1 Score and execution time. The proposed algorithm excels in all parameters.

Key Words: Information Retrieval, Machine Learning, Decision Tree, Random Forest, Naïve Bayes

Introduction

Access to situational information is essential during a catastrophe [1] as it aids in minimizing deaths, averting additional disasters, minimizing economic losses, and minimizing civil unrest [2]. Recent years have seen an increase in the use of microblogging platforms like Twitter as informational tools in disaster management [3] [4]. Twitter contains several words related to disasters that needs to process to identify current situation. Several research have been done to retrieve useful and valid information from twitters [5]. Twitter contains information in text form so several text classification techniques [6] [7] [8] are vital for this task. Some popular machine learning algorithms such as decision tree, random forest and Naïve Bayes are used extensively in text classification related work on social media platform like twitters [9] [10] [11] [12] [13]. Compared to other algorithms, decision trees need less data cleansing and are effective for situations involving decisions. However, it contains a number of restrictions regarding the classification:

- It could have a problem with overfitting.
- The computational complexity of the decision tree may rise with additional class labels.
- Accuracy and result generation might be affected.

Another popular and important classification algorithm is random forest. However, it has a number of shortcomings:

- There are more trees involved in random forest so training the model frequently takes more time and resources.
- As there are more trees involved, random forests sometimes require more time and space to train the model.
- Random forests may also overfit for noisy data.

Another well-known approach for classifying texts is naive bayes. It effectively resolves multi-class prediction problems and is based on the Bayes theorem. However, it has the following drawbacks:

- It makes the supposition that each characteristic is independent. Although it may seem wonderful in principle, you seldom ever discover a group of independent qualities in practice.
- The Naive Bayes model will give it 0 probability and won't be able to predict anything if your test data set contains a categorical variable of a category that wasn't included in the training data set.

Word embedding [14] [15] [16] is a crucial technique for text categorization that enables us to extract information from the text. For effective text retrieval, these attributes can be used as input in machine learning models. Dense word vectors, such as Word2Vec [17] [18] [19] and GLoVE [20] [21], are compact representations of a word's semantic meaning. The two models are trained differently, resulting in word vectors with somewhat different characteristics. The foundation of the glove model is the use of global word-to-word co-occurrence counts throughout the whole corpus. On the other hand, Word2vec makes use of co-occurance in local context. Word2vec is popular and easy to implement so it is widely used in the research work. A feed-forward neural network-based model [22] called Word2Vec is used to find word embeddings. The skip-gram model [23]and the CBOW [24] model are the two models that are frequently used to train these embeddings. Each word in the corpus serves as the input for the Skip-gram model, which then sends each word to an embedding layer in a hidden layer where it predicts the surrounding words. After training, the word is fed as input to the model, and the hidden layer value is used as the final embedding vector to produce the embedding for that specific word. The Continuous Bag of Words (CBOW) model predicts the original word from the context words for the input word that are sent to a hidden layer (embedding layer). The embedding for a specific word is again created after training by using the word as input and the hidden layer value as the final embedding vector.

The next section of the paper will describe the glimpses of literature survey. After that it will describe the proposed algorithm . Then it will describe the dataset and results of experiment. At last it will provide the conclusion of the paper.

Glimpses of Literature Survey

Many studies described the text classification using social media data of twitter. Some popular machine learning algorithms are employed for effective text classification of these twitter data. However, several research gaps are indetified on the basis of literature survey, which are highlighted as under :

- Text classification accuracy is not satisfactory of popular algorithms perticularly on natural disaster related twitters.
- Time it takes to execute result is also needs to be minimized.
- Without applying word emdeding , the results are not satisfatory.
- Decision of selecting appropriate word embedding technique is also challanging task.

The proposed research addresses the above challenges and design one algorithm to increase accuracy and mimimize execution time of result generation.

Proposed Word Embedding based Algorithm for text classification

The Table-1 describes the steps of proposed Algorithm.

Table-1 Proposed Word Embedding based Algorithm

Step-1 : Take Inputs as Number of Tweets
T = { t1,t2,t3,…tn}
Where t1….tn number of tweets

Step-2 : Pre-processing of Tweets

(stop word removal, QE, Stemming)

Step-3 : Apply Word Embedding

Step-4 : Apply Text Classification

$$X^a y + b = 0$$

Where X= Vector of Tweets, b= distance, y = data point

$X^{\wedge}$ = 1 or 0 ; 1 if $X^a y + b \geq 0$ else 0

Step-5 : Result Generation

(Parameter used: Precision, Recall, F1, Execution Time)

Step-6 : Evaluation of Results

Step-7 : Repeat Step-4, if necessary, until to achieve desire accuracy

Dataset and Results of Experiments

The Microblog Track at The Text Retrieval Conference Supports the evaluation of microblog information retrieval, and it is among the first in the IR community to give a common assessment of information retrieval systems. Twitter and other microblogging platforms are increasingly being utilized to boost relief efforts when a disaster strikes.Datasets for numerous experiments are provided via the IRMiDis track in FIRE. Annotators are given tweets to categorize into many categories, such as need tweets, availability tweets, fact-checkable and non-fact-checkable tweets, to aid with disaster efforts. The researchers take part and submit their runs, which are evaluation findings. The dataset for proposed reserch is collected from Forum for Information Retrival Evaluation (FIRE). The dataset comprises of more than 50,000 tweets of Nepal earthquake. The FIRE community's microblog dataset included a selection of tweets about the April 25, 2015 earthquake in Nepal. It was observed

that many Twitter users began sharing tweets both during and after the earthquake.The tweets discussed the earthquake's events and the relief efforts.However, because they were being retweeted, numerous tweets included the same information. The most crucial step in creating the microblog collection was removing duplicate tweets because they can cause the IR system's performance to be overestimated and may also increase the workload for the human annotators by creating information overload.The sample dataset is described in Table-2.

<num> **Number:T1**<title> **What resources were available** <desc> Identify the messages which describe the availability of some resources. <narr> A relevant message must mention the availability of some resource like food, drinking water, shelter, clothes, blankets, blood, human resources like volunteers, resources to build or support infrastructure, like tents, water filter, power supply, etc. Messages informing the availability of transport vehicles for assisting the resource distribution process would also be relevant. Also, messages indicating any services like free wifi, sms, calling facility etc. will also be relevant. However, generalized statements without reference to any particular resource or messaging asking for donation or money would not be relevant.
<num> **Number: T2**<title> **What resources were required** <desc> Identify the messages which describe the requirement or need of some resources. <narr> A relevant message must mention the requirement / need of some resource like food, water, shelter, clothes, blankets, human resources like volunteers, resources to build or support Infrastructure likes tents, water filter, power supply, blood and so on. A message informing the requirement of transport vehicles assisting resource distribution process would also be relevant. Also, messages requesting for any services like free wifi, sms, calling facility etc. will also be relevant. However, generalized statements without reference to any particular resource or messaging asking for donation or money would not be relevant.
<num> **Number:**T3<title>What medical resources were available <desc> Identify the messages which gives information about availability of medicines & other medical resources. <narr>A relevant message must mention the availability of some medical resource like medicines, medical equipments, blood, supplementary food items(e.g, milk for infants), human resource like doctors/staff and resources to build or support medical infrastructure like tents, water filters, power supply, ambulance etc. generalized statements without reference to medical resources would not be relevant.
<num> **Number:**T4<title> What medical resources were required <desc>Identify the messages which describe the requirement of some medicines & other medical resources <narr> A relevant message must mention the requirement of some medical resource like medicines, medical equipments, supplementary food items, blood , human resource like doctors/staff and resources to build or support medical infrastructure like tents, water filters, power supply, ambulance etc. generalized statements without reference to medical resources would not be relevant.
<num> **Number:**T5<title> What infrastructure damage and restoration were being reported <desc> Identify the messages which contain information related to infrastructure damage or restoration. <narr> A relevant message must mention the damage or restoration of some specific infrastructure resources, such as structures(e.g. roads, runways,railway), electricity, mobile/Internet connectivity, etc. general statements without reference to infrastructure resources would not be relevant.

Table-2 Sample Dataset

Proposed reserch used a rule-based pre-processor to get rid of stopwords and undesirable symbols because the tweets were filled with extraneous punctuation, URLs, special symbols,

and emoticons. Later on, the conventional Porter stemmer was used to stem the tweets as part of the pre-processing stage. The Word2Vec model generates a vector for each term in the corpus, and we can identify the relevant terms using vectors that fall in the bracket of cosine similarity. The proposed research used the Word2Vec technique for tweets. After proper preprocessing and word embedding , proposed text classification algorithm is applied on the data and also compared the proposed technique with popular machine learning algorithms such as Decision Tree, Random Forest and Naïve Bayes. The precision comparison of proposed algorithm and other well-known algorithms with and without word embedding is described in the table-3. The result is derived in macro and micro measurement.

Micro-precision= (True Positive1 + True Positive2)/ (True Positive1 + False Positive1 +

True Positive2+ False Positive2 ----------(1)

Macro-precision= (Precison1 + precision2)/2 ----------(2)

Table-3 Precision comparison of Proposed Algorithm and other Algorithms

Algorithm	**Precision**			
	Without Word Embedding		**With Word Embedding**	
	Macro	**Micro**	**Macro**	**Micro**
Decision Tree	0.63	0.66	0.66	0.70
Naive_Bayes	0.49	0.52	0.52	0.58
Random Forest	0.78	0.80	0.82	0.84
Proposed Algorithm	0.75	0.77	0.80	0.85

The table-3 result described that proposed algorithm precision is similar to Random Forest Algorithm. It is far better than all other algorithms. The main limitation of the random forest is there are more trees involved, random forests sometimes require more time and space to train the model.

The recall comparison of proposed algorithm and other well-known algorithms with and without word embedding is described in the table-4. The result is derived in macro and micro measurement.

Micro-recall= (True Positive1 + True Positive2)/ (True Positive1 + False Negative1 + True Positive2+ False Negative2) ----------(3)

Macro-recall= (Recall1 + Recall2)/2 ----------(4)

Table-4 Recall comparison of Proposed Algorithm and other Algorithms

Algorithm	Recall			
	Without Word Embedding		With Word Embedding	
	Macro	Micro	Macro	Micro
Decision Tree	0.59	0.62	0.62	0.65
Naive_Bayes	0.48	0.55	0.50	0.58
Random Forest	0.39	0.44	0.42	0.47
Proposed Algorithm	0.62	0.68	0.66	0.71

The table-4 result described that proposed algorithm recall value is far better than all other algorithms.

The F1 Score comparison of proposed algorithm and other well-known algorithms with and without word embedding is described in the table-5. The result is derived in macro and micro measurement.

Micro-F1= (Micro precision. Micro recall)/ (Micro precision+Micro recall) ----------(5)

Macro-F1= (Macro precision. Macro recall)/ (Macro precision+Macro recall)----------(6)

Table-5 F1 Score comparison of Proposed Algorithm and other Algorithms

Algorithm	F1			
	Without Word Embedding		With Word Embedding	
	Macro	Micro	Macro	Micro
Decision Tree	0.61	0.64	0.62	0.67
Naive_Bayes	0.48	0.53	0.48	0.58
Random Forest	0.49	0.57	0.55	0.60
Proposed Algorithm	0.67	0.72	0.7	0.77

The table-5 result described that proposed algorithm F1 score is far better than all other algorithms. Table-6 describes the execution time of all algorithms in seconds.

Table-6 Execution Time of all algorithms

Algorithm	**Algorithm Execution Time**
Decision Tree	0.0249 seconds
Naïve_Bayes	0.0667 seconds
Random Forest	0.0229 seconds
Proposed Algorithm	0.0059 seconds

The result described that the execution time of proposed algorithm is mimimum than all other algorithms. Overall, the proposed algorithm is far better in all parameters i.e precision, recall, F1-Score and Execution Time.

Conclusion

The main objective of this reserch paper was to design novel algorithm using word embedded technique to improve accuracy and reduce execution time for information retrival using social media platform twitter. To conduct the microblog retrieval challenge, the proposed algorithm conducted two separate experiments. It was observed that the proposed algorithm with word emdedding technique performs fairly well in the classification test, and also produced a respectable accuracy score by employing this approach. The proposed algorithm precision was similar to Random Forest Algorithm but random forest sometimes requires more time and space to train the model. As far as recall, F1-score and Execution time is concerned, the proposed algorithm excel than all other algorithms. Overall proposed algorithm performed better in all dimensions than other algorithms.

References

[1] "World Disasters Report 2013 - Focus on technology and the future of humanitarian action," http://www.ifrc.org/PageFiles/134658/WDR 2013 complete.pdf, 2013.

[2] M. B. S. G. S. Basu, "Post disaster situation awareness and decision support through interactive crowdsourcing," *In: Proc. International Conference on Humanitarian Technology: Science, Systems and Global Impact (HumTech). Procedia Engineering,,* p. 167–173, 2016.

[3] M. C. C. D. F. V. S. Imran, "Processing Social Media Messages in Mass Emergency: A Survey," *ACM Computing Surveys,* vol. 47, no. 4, pp. 1-67, 2015.

[4] I. e. a. Varga, "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* , 2013.

[5] "AIDR - Artificial Intelligence for Disaster Response," https://irevolutions.org/2013/10/01/aidrartificial-for-disaster-response.

[6] V. M. Kanimozhi KV, "Unstructured data analysis-a survey," *Int J Adv Res Comput Commun Eng,* vol. 4, no. 3, pp. 223-225, 2015.

[7] M. e. a. Grčar, "Stance and influence of Twitter users regarding the Brexit referendum," *Computational social networks,* vol. 4, pp. 1-25, 2017.

[8] W. e. a. Fan, "Tapping the power of text mining," *Communications of the ACM,* vol. 49, no. 9, pp. 76-82, 2006.

[9] A. I. P. a. A. M. T. Ashari, "Performance comparison between Naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool," *International Journal of Advanced Computer Science and Applications,* vol. 4, no. 11, 2013.

[10] W. B. e. a. Zulfikar, "The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter," in *017 5th International Conference on Cyber and IT Service Management (CITSM). IEEE*, 2017.

[11] A. I. P. a. A. M. T. Ashari, "Performance comparison between Naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool," *International Journal of Advanced Computer Science and Applications,* vol. 4, no. 11, 2013.

[12] A. a. A. K. Rane, "Sentiment classification system of twitter data for US airline service analysis," in *IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, 2018.

[13] S. S. A. a. M. S. Sharma, "Degree based classification of harmful speech using twitter data," *arXiv preprint arXiv,* vol. 1806, no. 04197, 2018.

[14] S. e. a. Lai, "How to generate a good word embedding," *IEEE Intelligent Systems,* vol. 31, no. 6, pp. 5-14, 2016.

[15] B. e. a. Wang, "Evaluating word embedding models: Methods and experimental results," *APSIPA transactions on signal and information processing,* vol. 8, 2019.

[16] S. e. a. Ghannay, "Word embedding evaluation and combination," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2016.

[17] K. W. Church, "Word2Vec," *Natural Language Engineering ,* vol. 23, no. 1, pp. 155-162, 2017.

[18] X. Rong, "word2vec parameter learning explained," *arXiv preprint arXiv,* vol. 1411, no. 2738, 2014.

[19] G. A. B. a. F. A. P. Di Gennaro, "Considerations about learning Word2Vec," *The Journal of Supercomputing,* pp. 1-16, 2021.

[20] J. R. S. a. C. D. M. Pennington, "Glove: Global vectors for word representation," in *Proceedings of*

the 2014 conference on empirical methods in natural language processing (EMNLP), 2014.

[21] E. M. e. a. Dharma, "The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification," *J Theor Appl Inf Technol,* vol. 100, no. 2, 2022.

[22] M. H. Sazli, "A brief review of feed-forward neural networks," *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering,* vol. 50, no. 01, 2006.

[23] C. McCormick, "Word2vec tutorial-the skip-gram model," http://mccormickml. com/2016/04/19/word2vec-tutorial-the-skip-gram-model, 2016.

[24] B. Liu, "Text sentiment analysis based on CBOW model and deep learning in big data environment," *Journal of ambient intelligence and humanized computing,* vol. 11, pp. 451-458, 2020.

[25] M. H. Sazli, "A brief review of feed-forward neural networks," *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering,* vol. 50, no. 01, 2006.