Acta Scientific COMPUTER SCIENCES

Volume 5 Issue 3 March 2023

Research Article

Classification of Disaster Tweets Using Natural Language Processing Pipeline

S Deepa Lakshmi[1] and T Velmurugan[2]*

[1]Assistant Professor, PG and Research Department of Computer Science, Dwaraka Doss Goverdhan Doss Vaishnav College, Chennai, India

[2]Associate Professor, PG and Research Department of Computer Science, Dwaraka Doss Goverdhan Doss Vaishnav College, Chennai, India

***Corresponding Author:** T Velmurugan, Associate Professor, PG and Research Department of Computer Science, Dwaraka Doss Goverdhan Doss Vaishnav College, Chennai, India.

Received: January 31, 2023

Published: February 28, 2023

Abstract

A number of methods are utilised for the analysis of tweets based information extraction. Natural Language Processing (NLP) is a branch of artificial intelligence that enables us to understand human sentences and words. NLP combines rule-based modelling of human language combined with statistical, machine learning and deep learning models. This research work aims at using NLP for disaster tweet classification using pipelines. Tweets are highly unstructured in nature and hence text pre-processing is an important phase which involves removing unwanted and irrelevant words from the tweets. NLP pipeline is a set of steps followed to build end to end NLP software including text pre-processing, feature extraction and modelling. Pre-processing is done using tokenization, stop words removal, lemmatization and feature extraction using TF-IDF transformer. To analyse the tweets based informations, classification algorithms are used. The classification algorithms Support Vector Machine, MLP, Adaboost and Multinomial NB are used to classify the tweets and the best performing classifier is identified.

Keywords: Natural Language Processing Pipeline; Feature Extraction; Classification of Tweets; Multinomial NB

Introduction

Text mining or Text Analytics is an artificial intelligence technology that uses NLP to convert the raw text in human readable form into structured data suitable for analysis using machine learning techniques [1]. The structured data can be used for descriptive and predictive analytics. NLP uses techniques to interpret vagueness in human language like automatic summarization, parts-of-speech tagging, entity extraction, NLP understanding and recognition. NLP finds application in many areas including email spam/ham filtering, smart assistant, language translation, search engines, text analytics etc. [2].

Social media platforms are a common choice for people to express their feelings and the amount of data generated is enormous. This provides an insight into the sentimental reaction of people using various data analytics tools and algorithms. Such analysis using advanced machine learning algorithms can be utilized by emergency/disaster management teams [3]. The challenging part here is the identification of tweets that indicate information pertaining to disaster. Hence it is vital to develop a solution to enhance the ability of machine learning algorithms. The aim of this work is to employ data analytic techniques to build a classification model that can accurately identify tweets referring to real disasters. The Organization of the paper is as follows: section 2 describes related work done by researchers, section 3 explains the architecture of NLP pipeline, section 4 discusses the result analysis and finally section 5 concludes the findings of the research work.

Related work

The use of data from social networks has increased in recent years for sentiment analysis, political campaigns, product rating etc. Many researchers have investigated using various data analytic tools and techniques. A voting classifier was proposed for sentiment analysis which is based on logistic regression and stochastic descent classifier in 2019 by a researcher Rustam., *et al*. [4]. Domain specific seed list was used to build user profile using twitter data which helps in providing personalized recommendations in 2020 [5]. A web based application was developed to classify tweets into four categories of topics in 2016 by researcher Indra., *et al*. Tweets are fetched, pre-processed, feature extracted and machine learning techniques applied [6]. Stemming is used popularly by researchers and a researcher Ahmad., *et al*. in 2016 analysed the influence of stemming on tweet classification [7]. Recently in 2022, users sentiments were analysed using techniques such as tfidf, word2vec, glove and fast text to obtain feature subsets [8] and Classifying sentiments was done using lexicon based approach where seven classifiers were used to classify the tweets [9]. In 2023, a twitter based disaster response system that uses recurrent nets for training the classifier was used [10].

NLP pipeline

Twitter has become an important communication media in times of emergency and as a tool for recreation. People are more inclined to use their mobile phones during emergency situations to seek help. Twitter has become a popular choice and disaster relief organizations are interested in monitoring twitter messages. Tweets are not always clear in terms of the context. The challenge is to identify the tweets that are about the real disaster and which are not.

Data set description

The dataset was created by figure-eight company that has 10000 tweets that were hand classified and the dataset was obtained from kaggle competition run by kaggle.com. The dataset contains a unique identifier for each tweet, the text of the tweet, the location of the tweet, a keyword from the tweet and target denoting whether the tweet is about a real disaster or not. If the tweet is about a real disaster, it is assigned a value 1 and 0 otherwise.

Data pre-processing

The tweets are not structured and need to be pre-processed. Figure 1 shows the architecture of the NLP pipeline used in this research work. This research work uses python for exploratory data analysis implemented in jupyter notebook. The text of the tweet is tokenized by a user defined function tokenize that pre-process the text data. The entire text is converted to lowercase and special characters, symbols and mail ids are removed. The text is then split into words using natural language toolkit library (NLTK). This is an important step as words are required for further analysis. Stop words are words that occur commonly across all the documents and do not add much information to the text. Such stop words are removed from the text using stop word removal available in NLTK. Lemmatization is the process of grouping words together so as to analyse them as a single item [11]. It links words with similar meanings to one word. Lemmatization is applied to the text data and clean tokens are obtained.

Feature extraction is the process of transforming raw data into numerical features while preserving the information in the original dataset. Count Vectorizer is used to convert text into a matrix of word counts called document term matrix where terms are represented as columns and documents as rows known as bag of words [12]. TfidfTransformer is the feature extraction method used. Tfidf means term-frequency times inverse document-frequency [13].

IDF= log[(1+D) / (1+df(d,t))] + 1

Where D is the number of documents and df(d,t) is the number of documents a term has appeared in the document term matrix. The Tf-idf scores are computed using the above formula.

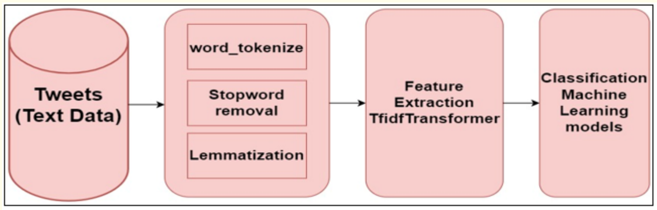
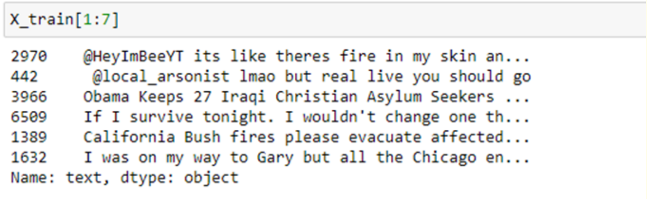

Tweets
(Text Data)
word_tokenize
Stopword removal
Lemmatization
Feature Extraction TfidfTransformer
Classification Machine Learning models

Figure 1: Pipeline Architecture.

The pre-processed text data is classified using various classifiers and the performance of the classifiers is compared in this work. The classifiers used in this research work are support vector machines, multilayer perceptron, adaboost, multinomial NB and random forest. The parameters of the classifiers are modified to obtain best performance. The performances of the classifiers are measured using precision, f-score and accuracy.

Results and Discussion

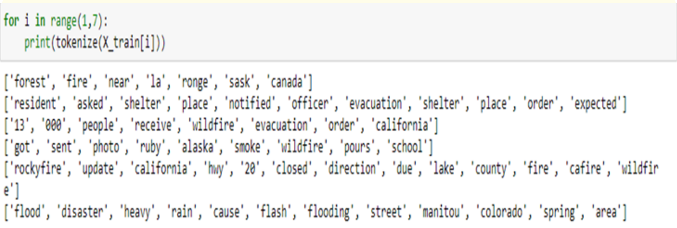
The tweets are pre-processed by converting to lower case letters, substituting special characters, punctuations by empty string using regular expressions. The text is then converted to words by word_tokenize function available in python. Stop words are then removed from the bag of words. Words with similar meaning are grouped together using lemmatization and clean tokens are obtained. A sample of the tokens obtained is given below. Figure 2 shows the raw text data of the tweets and figure 3 shows the pre-processed data.



```
X_train[1:7]

2970    @HeyImBeeYT its like theres fire in my skin an...
442      @local_arsonist lmao but real live you should go
3966    Obama Keeps 27 Iraqi Christian Asylum Seekers ...
6509    If I survive tonight. I wouldn't change one th...
1389    California Bush fires please evacuate affected...
1632    I was on my way to Gary but all the Chicago en...
Name: text, dtype: object
```

Figure 2: Raw Tweets data.



```
for i in range(1,7):
    print(tokenize(X_train[i]))

['forest', 'fire', 'near', 'la', 'ronge', 'sask', 'canada']
['resident', 'asked', 'shelter', 'place', 'notified', 'officer', 'evacuation', 'shelter', 'place', 'order', 'expected']
['13', '000', 'people', 'receive', 'wildfire', 'evacuation', 'order', 'california']
['got', 'sent', 'photo', 'ruby', 'alaska', 'smoke', 'wildfire', 'pours', 'school']
['rockyfire', 'update', 'california', 'hwy', '20', 'closed', 'direction', 'due', 'lake', 'county', 'fire', 'cafire', 'wildfir
e']
['flood', 'disaster', 'heavy', 'rain', 'cause', 'flash', 'flooding', 'street', 'manitou', 'colorado', 'spring', 'area']
```

Figure 3: Pre-processed Tweets.

A countvectorizer is used to count the number of words in the pre-processed tweets and idf values are computed. The lower the idf value of a word, the less unique it is to any particular document. The idf values for some of the words found in a particular tweet is given in the table 1. The words grief, broken, bioterrorism etc have the same idf values indicating the importance of such words in classifying a tweet as related to disaster.

Table 1: Result of idf values for the tweets.

Tweets	idf_weights	Tweets	idf_weights
Grief	6.942799	Bit	6.942799
Broken	6.942799	Crashes	6.942799
Planned	6.942799	Smithsonian	6.942799
Wrong	6.942799	Secret	6.942799
Bioterrorism	6.942799	British	6.942799

The tfidf scores are computed and the more common the words across documents, the lower its score. The more unique a word is to a document, the higher the score. The table 2 shows the tfidf values of the words found in a tweet. The tfidf values always scale between 0 and 1. The higher the value, more relevant is the word to the document. The word 'wrap', 'female' are more relevant to the document than the word 'news'.

Table 2: Values of tfidf for the tweets.

Tweets	tfidf weights	Tweets	tfidf weights
Wrap	0.352813	Afghanistan	0.318118
Female	0.341033	Un	0.318118
Warns	0.331896	Rise	0.275129
Iraq	0.32443	Child	0.273131
Casualties	0.26284	News	0.192895

The pre-processed text data with tfidf values are classified as text pertaining to disaster or not using the classifiers mentioned above. Table 3 shows the results of classification obtained using the classifier Multinomial NB.

The text in table 3 refers to the tweets obtained from twitter and the target refers to the classification as disaster or not classified by the classifier. In table 3, the tweets pertaining to natural disaster are classified with a value 1 and the tweets not pertaining to disaster are classified with a value 0. The tweet," there is a forest fire at spot pond, geese are fleeing across the street, I cannot save them all", is classified as disaster with a value 1 whereas the tweet," Who is bringing the tornadoes and floods. Who is bringing the climate

Table 3: Results of classification using Multinomial NB.

Text	Target
And Kolkata is struck by a Cyclonic Storm. Sumthng big is gonna happen 2day evng. Heavy rains nd a violent storm approachng. God help us.	1
Heard about #earthquake is different cities, stay safe everyone.	0
there is a forest fire at spot pond, geese are fleeing across the street, I cannot save them all	1
Apocalypse lighting. #Spokane #wildfires	1
Hey! How are you?	0
We're shaking...It's an earthquake	1
Why is it that my pinky feels like it's lit on fire ? #freaky	0
Who is bringing the tornadoes and floods. Who is bringing the climate change. God is after America He is plaguing her	0

change. God is after America He is plaguing her" is classified as not a disaster. The tweet, "Heard about #earthquake is different cities, stay safe everyone" discusses about earthquake which is a natural disaster and it is not classified as disaster. Though the words fire, flood, earthquake and tornadoes are specified, it is not classified as disaster as the tweet discusses some disaster but it does not pertain to a disaster requiring help from disaster management team.

The performance of the classifiers in terms of the precision, recall and accuracy are discussed as shown in figure 4. Accuracy is how close a measured value is to the true value [14]. SVC has the highest accuracy followed by MLP, Multinomial NB and Adaboost. The parameters of the classifiers are tuned to improve the performance of the classifiers.

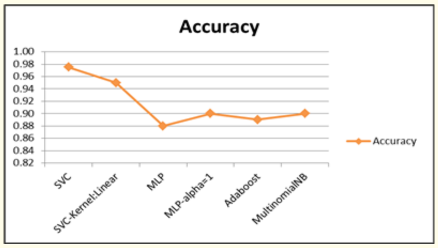


Figure 4: Accuracy measures of the classifiers.

The accuracy of support vector machine is high compared to all other classifiers and Multilayer perceptron classifier has the least accuracy as evident from figure 5. Fscore is the measure of a test's accuracy [15] and Adaboost has the lowest fscore value whereas support vector machine, multilayer perceptron have high values of fscore.

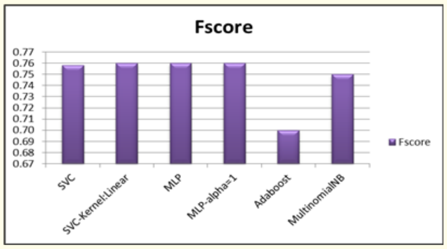


Figure 5: F Score measures of the classifiers.

Figure 6 compares the precision obtained by the classifiers. Precision is the ability of a classification model to identify only the relevant data points[16]. Multinomial Naïve Bayes has a high precision value and followed by support vector machine and MLP. Adaboost has the lowest precision value.

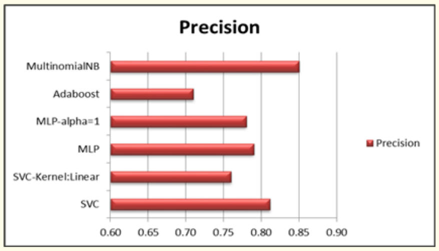


Figure 6: Precision measures of the classifiers..

Considering all the metrics, Adaboost has the least performance and SVC has better performance in terms of accuracy and fscore. Multinomial NB has high performance in terms of precision, fscore but has lower accuracy. MLP also has high precision and fscore value and low accuracy. Considering the metrics collectively, Support vector machine has the best performance in terms of accuracy, fscore and precision [16].

Conclusion

Human language understating and generation is a popular technique in the current world applications for text processing. NLP techniques using machine language enables to understand and decode human language. This research work focuses on using NLP techniques to classify tweets during a disaster which is used by the disaster management team. An NLP pipeline is used to preprocess the text data containing the tweets removing irrelevant and unwanted text and converting to a bag of words. It is then converted into numerical values using feature extraction techniques. The extracted features are used by classifiers to classify the tweets as containing information pertaining to real disaster or not. The classifier support vector machine had the best performance in classifying the tweets in terms of precision, fscore and accuracy compared with the other methods.

Bibliography

1. Berry MW and Kogan J. “Text mining: applications and theory”. John Wiley and Sons (2010).

2. Rastenis J., *et al*. “Multi-language spam/phishing classification by email body text: Toward automated security incident investigation”. *Electronics* 10.6 (2021): 668.

3. Balogun AL., *et al*. “Assessing the potentials of digitalization as a tool for climate change adaptation and sustainable development in urban centres”. *Sustainable Cities and Society* 53 (2020): 101888.

4. Rustam Furqan., *et al*. “Tweets classification on the base of sentiments for US airline companies”. *Entropy* 21.11 (2019): 1078.

5. Khattak AM., *et al*. “Tweets classification and sentiment analysis for personalized tweets recommendation”. *Complexity* (2020).

6. Indra ST., *et al*. “Using logistic regression method to classify tweets into the selected topics”. In 2016 international conference on advanced computer science and information systems (icacsis)”. (2016): 385-390.

7. Hidayatullah AF., *et al*. “Analysis of stemming influence on indonesian tweet classification”. TELKOMNIKA (Telecommunication Computing Electronics and Control) 14.2 (2016): 665-673.

8. Didi Yosra., *et al*. “COVID-19 Tweets Classification Based on a Hybrid Word Embedding Method”. *Big Data and Cognitive Computing* 6.2 (2022): 58.

9. Gulati K., *et al*. “Comparative analysis of machine learning-based classification models using sentiment classification of tweets related to COVID-19 pandemic”. *Materials Today: Proceedings* 51 (2022): 38-41.

10. Lamsal R and Kumar TV. “Twitter-based disaster response using recurrent nets”. In Research Anthology on Managing Crisis and Risk Communications (2023): 613-632.

11. Korenius T., *et al*. “Stemming and lemmatization in the clustering of Finnish text documents”. In Proceedings of the thirteenth ACM international conference on Information and knowledge management (2004): 625-633.

12. Kulkarni A., *et al*. “Converting text to features. Natural Language Processing Recipes”. Unlocking Text Data with Machine Learning and Deep Learning Using Python (2021): 63-106.

13. Zhao G., *et al*. “TFIDF based feature words extraction and topic modeling for short text”. In Proceedings of the 2018 2Nd International Conference on Management Engineering, Software Engineering and Service Sciences (20180): 188-191.

14. Osisanwo FY., *et al*. “Supervised machine learning algorithms: classification and comparison”. *International Journal of Computer Trends and Technology (IJCTT)* 48.3 (2017): 128-138.

15. El Rahman SA., *et al*. “Sentiment analysis of twitter data”. In 2019 international conference on computer and information sciences (ICCIS) (2019): 1-4.

16. Davis J and Goadrich M. “The relationship between Precision-Recall and ROC curves”. In Proceedings of the 23rd international conference on Machine learning (2006): 233-240.