ISSN 2278-3091
Volume 9, No.1.1, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse0691.12020.pdf
https://doi.org/10.30534/ijatcse/2020/0691.12020

Applying Modified TF-IDF with Collocation in Classifying Disaster-Related Tweets

Gleen A. Dalaorao[1], Ariel M. Sison[2], Ruji P. Medina[3]
[1]Technological Institute of the Philippines, Philippines, gadalaorao@carsu.edu.ph
[2]Emilio Aguinaldo College, Philippines, ariel.sison@eac.edu.ph
[3]Technological Institute of the Philippines, Philippines, ruji.medina@tip.edu.ph

ABSTRACT

Disaster-related tweets classification refers to posted tweets on twitter during the time-critical events (e.g., natural, human-made) that are group together according to pre-defined categories (e.g., donations, awareness, help, etc.). The purpose of classification is to deliver the conveyed messages to suitable authorities on time for requiring needs for immediate action. The classification works well if terms that are extracted from tweets are carefully selected or referred to as good attributes that can best label the uncategorized tweets. The Term Frequency – Inverse Document Frequency (TF-IDF) helps classifier to extract useful features. Hence, TF-IDF naturally works on distinct terms. However, a single term occasionally can be ambiguous, which means that when a separate term used for indexing, could carry numerous connotations. The distinct term can sometimes too broad, which means it does not have a discerning power to discriminate terms, for example, from the two individual terms "college" and "junior". These two terms are not adequate to differentiate “college junior” from "junior college" [1]. Hence, applying traditional TF-IDF in text classification can reduce classification efficacy. Thus, a combination of terms known as collocation is introduced as an improvement to TF-IDF to boost the text classification effectiveness. This paper aims to provide an analysis of the efficacy of tweets classification by applying improved TF-IDF with collocation. This experiment utilizes tweets dataset from the CrisisLex website. The performance evaluation metrics considered are confusion matrix, precision, recall, and F1 score. The result shows that there is a favorable increase in the proposed study as compared to traditional TF-IDF through said evaluation metrics vary from 4% to 24%. The study also establishes that RandomForest consistently outperforms that two other compared classifiers.

Key words : TF-IDF, Term Weighting, Collocation, Disaster Related Tweets, Tweet Classification, RandomForest, SVM, Multinomial Naïve Bayes.

1. INTRODUCTION

Twitter is known as a useful tool to share information, collaborate, and communicate with other people [1]. There is a lot of literature quoting that twitter is utilized during time-critical events (e.g., natural, human-made) such as disseminating weather-related announcements, tweeting for help or donations, etc. The type of tweets that are used during the time-critical events is tag as disaster-related tweets [2]. So, classifying tweets on time is essential so that it can deliver the conveyed messages to suitable authorities as for requiring needs for immediate action [3]. Thus, there is a need to take out terms which can represent the picture of the content of the tweets, so that it can be classified tweet according to its purpose in a short time [4][5]. The TF-IDF supports classifiers to take out the appropriate feature terms. The TF-IDF further clarifies that it is projected to mirror the significance of a particular term in a specified document. The term “relevance” means it’s comparative to the number of facts that it provides about its perspective, regardless if it is a text, document collection, or a corpus [6]. The most important terms are those terms that give a hint to humanity for what the entire document is all about [7]. The TF-IDF performs by assigning a value on each term in the document. These weighted terms are reflected in the matrix. The TF-IDF assumes that when a term seems less in document collections, then the term is significant, and the term has a higher TF-IDF weight. However, if a term is visible on most of the document collection, then this term is known to be as typical. These terms have a lesser TF-IDF weight [8].

However, traditional TF-IDF weighted the terms independently. On the contrary, when single terms are utilized as the index can sometimes misleading, as an illustration, for the compound word 'super typhoon' could have several interpretations, which might lead to slip-up when utilized as different indexing. The distinct term is also too broad, which means that a single term does not have a discerning power to distinguish terms, for instance, as for the two individual terms such as "junior" and "college." These individual terms is not adequate to discriminate "college junior” from "junior college" [8][9]. As a consequence, the incorporation of a grouping of terms or collocation is more descriptive and desirable by humans. In this fashion, this helps to enriched the classification effectiveness [9] [10][11][12]. The overall notion is that reducing the number of terms and keeping alone the relevant terms in the matrix can support increase the text classification effectiveness [13].

As a result, this study will provide an analysis of the effectiveness of improved TF-IDF with collocation as applied

in disaster-related tweets classification. The result of this study is further assessed using several evaluation metrics, namely, confusion matrix, precision, recall, and F1-score [14].

2. RELATED LITERATURE

The following related kinds of literature are the basis of this study.

2.1 Twitter

Twitter is a micro-blog where people can openly communicate with each other on a vast array of topics, and this capacity has led Twitter to be one of the world's biggest social media sites. The twitter has been utilized during the time of time-critical events (.e.g. natural, human-made) such as disseminating weather-related pronouncements such as tweeting for help, donations, etc. [15].

2.2 The Traditional TF-IDF

The research paper of [12] presents the N-Gram model as a resolution of the shortcoming of the Bag-of-Words (BOW) method. The article clarifies that BOW neglects the relationship among terms. The N-gram model is a method used to find some sort of item in the text, on which it considers grouping or association of terms. The study of [16], enhances the part of TF-IDF equation specifically on the IDF portion, in which the survey said that TF-IDF does not consider phrases or a combination of words. The study further emphasizes that there are terms that should be treated as one, such as Osaka University, Google Search, etc. The study reveals that the N-Gram model is used to retrieve compound words. The paper of [17] describes collocation as a group of terms that goes together naturally, which further elaborates that the full control of a language is gone afar from its semantic meaning and form of its single term. Its habitual word combination can fully understand the language. The work of [21] concentrates on improving TF-IDF by integrating collocation on the pre-processing step on the TF-IDF process The result of the cleaned dataset is analyzed by determining the terms with a specific pattern adjective + noun, noun + noun, noun + verb, etc. or any noun surrounding terms. The extraction process will start first on identifying the part of speech (POS) of each term before it determines if there exists a collocation pattern. The extracted collocation will be merged with a hyphen ("_") in between different unique terms to treat it as a single term. The result of collocation will be stored in a list together with non-collocate terms while disregarding stopwords. The result of the study is measured through precision, recall, and F1 score by comparing the result of traditional TF-IDF as compared to modified TF-IDF. The traditional TF-IDF functions as the standard since the traditional TF-IDF carry out the single term weight computation.

3. METHODOLOGY

3.1 Data Sources and Collections

This study has utilized the annotated datasets from the Crisislex website. These datasets are used in different researches. The Crisislex datasets contain various disaster-related tweets happening all of the worlds, such as Typhoon Haiyan, Boston Bombing, etc. [19][20]. This study doesn't constrain as to the number of target labels as applied in the experiment. The dataset is split into a 50% - 50% ratio, as applied to the train and testing set. The entire simulation is implemented in Anaconda Jupyter Python Notebook. The dataset contains various tweets on different event types such as earthquakes, floods, typhoons, etc. These tweets are labeled into multiple informative classes (e.g., affected Individuals, donation and volunteering, infrastructure and utilities, sympathy, and support). Additionally, as to overcome the problem of imbalance dataset or problem on under-sampling, since the majority of the disaster-related events share common attributes or labels, the proponent merges some of the events to increase the number of sampled tweets.

A. Overall Modified TF-IDF Classification Framework

The diagram, as shown in figure 1, illustrates that after tweets are extracted from the Crisislex dataset, the tweets are cleaned up first to ensure that tweets are free from noises such as punctuations, symbols, numbers, and non-English terms. The cleaned tweets will go to the modified TF-IDF process for the extraction of collocated and non-collocated terms and TF-IDF weight computation. The result of TF-IDF computation will be subjected to classification to determine the suitable document term category.

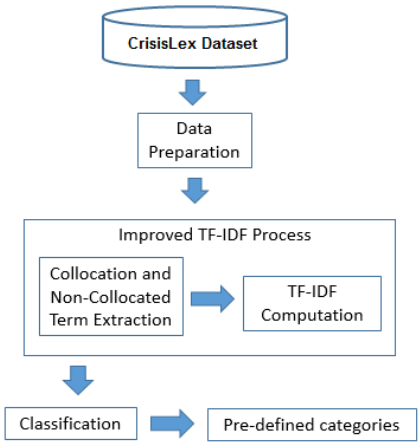


Figure 1: Overall Modified TFIDF Classification Framework

B. Data Preparation

This phase executes the data preparation where tweets undergo standard data cleaning techniques such as tokenization, lowercase conversion, removal of punctuations, numbers, and URL. This phase also includes the removal of non-English language. This phase ensures that tweets are cleaned or free from noise and ready for analysis. The outcome of data preparation will go through the modified TF-IDF process to extract the significant terms for classification, in which part of it is the collocation extraction. However, stop word removal is done after collocation has already been extracted. The POS tagging is applied to tweets to identify the specified collocation pattern. The considered collocations are those groups of terms that form a noun surrounding terms (adjective + noun, noun + noun, noun + verb, etc.). Additionally, there is a data cleaning technique

that is uniquely employed in this study, which excludes terms that are less than four (4) characters in length because these terms don't contribute much information. These terms cannot merely be filtered using stopwords because these terms do not belong to the stopwords set.

3.2 Term Collocation and Non-Collocation Extraction

The collocation has been described as a mixture of terms that goes together naturally, such as a beautiful girl, super typhoon, etc. [11]. The result of the cleaned dataset is analyzed by determining the terms which follow a specific pattern such as adjective + noun, noun + noun, noun + verb, etc. In the example, 'The pretty girl loves to eat a red apple'. The extraction process starts on identifying the part-of-speech of each term before it identifies if there exists a collocation pattern. So based on the example, the two terms pretty and the girl will be extracted as collocation since it follows adjective + noun pattern. These two terms recognized as collocation will be merged with a hyphen ("_") in between. The outcome of the collocation extraction process will be stored in a list together with non-collocated terms. The sample resulting list would appear like this [‘pretty_girl’, ‘loves’, ‘eat’, ‘red_apple’]. The pseudo-code is shown in figure 2.

```
1. Initialize acceptable_types list to 'Adjectives', 'Nouns',
   'Proper Nouns'
2. Initialize second_types list to 'Nouns', 'Proper Nouns'
3. while EverySentence in the Corpus
4.   while currentTerm in EverySentence
5.     If currentTermPOS in acceptable_types and
         nextTermPOS in second_types  then
6.       Merge the currentTerm and nextTerm with a
         hyphen in between and add to CollocationList
7.     Else
8.       Add to currentTerm to CollocationList
9.     Endif
10. Compute TF-IDF Collocation Matrix
```

Figure 2: TF-IDF Collocation Extraction Process Pseudo-code

3.3 TF-IDF Collocation Weight Processing

The result of the collocation extraction process is reflected in the TF-IDF collocation matrix. The result is shown in figure 4. The terms in the result set (collocated and non-collocated terms) are given a weight based on the equation shown in figure 3. The notion of the weight is that if a term appears on a less document means that this term is a proper candidate term for indexing, and this will have a higher TF-IDF weight. Still, when it seems all across tweets, then this term is typical, or cannot be a right candidate term to represent the document.

$$\text{TF-IDF weight} = \text{TF} \times \text{IDF}$$

where:

$$\text{TF} = \frac{\text{Total No. of Terms T appear on Tweet}}{\text{Total No. of Terms in the Tweet}}$$

$$\text{IDF} = \log\left(\frac{\text{Total No. of Tweets}}{\text{Total No. of Tweets containing Term T}}\right)$$

Figure 3: TF-IDF Equation

3.4 Tweets Classification Algorithm

Tweet classification is the way of classifying tweets or categorize tweets according to tweet content. Tweet (Text) Classification is one of the primary tasks in Natural Language Processing (NLP) with broad applications such as sentiment analysis, topic labeling, spam detection, etc. [11]. There are three (3) known tweet (text) classifiers that are utilized to these experiments as applied to both traditional TF-IDF and modified TF-IDF approaches, namely RandomForest, Multinomial Naïve Bayes, and SVM.

3.5 Evaluation of Traditional TFIDF vs Modified TFIDF

The tweet dataset is divided into a 50% - 50% ratio for training and testing datasets. The performance of the text classifiers as tested to two TF-IDF approaches will be assessed based on precision, recall metrics, and F1 score, in which the following equations do these metrics.

Precision = TP ⁄ (TP + FP)
Recall = TP ⁄ (TP + FN)
F1 = 2 x ((Precision*Recall)/ (Precision+Recall))

where the following acronyms are read as follows: TP means true positive, FP means false positive, TN means true negative, and FN means false negative [14][19]. The traditional TF-IDF serves as the baseline for comparison since the traditional TF-IDF performs the single term weight computation. The same parameters, setting, and dataset are applied to both TF-IDF approaches.

4. RESULTS AND DISCUSSION

Table 1 shows the number of classes with the corresponding number of tweets per disaster-related event before and after data cleaning. However, to minimize the partialities of the imbalance dataset. All of the disaster-related events under the Crisislex dataset, as mentioned in table 1, are combined into a single dataset to have a large dataset since all of the disaster-related tweets event datasets share standard class labels or attributes. The methods for data cleaning are tokenization, lowercase conversion, punctuation removal, URL's and numbers removal, and non-English language exclusion. The stopword removal is done after the identification of part-of-speech or collocation extraction. The stopwords are frequently used words like 'the', 'ourselves’, ‘hers’, ‘between’, ‘yourself’, etc. These type of words that are less-informative because these words are common to most of the documents [20]. These English stopwords are already part of NLTK package, in which this can be accessed using python command “set (stopwords.words ('English'))."

Table 1: Dataset Before and After Data Cleaning

Datasets	No. of Classes	Data Cleaning	
		Before	After
Alberta Floods	6	913	900
Costa Rica Earthquake	6	866	321
Colorado Floods	6	943	926
Australia Bushfire	6	930	917

Bohol Earthquake	6	943	785
Boston Bombing	6	913	892
Typhoon Yolanda	6	924	881
West Texas Explosion	6	883	861
Merge DataSet Total	6	7315	6483

Figure 4 illustrates the distribution of tweet instances per category. These number of tweets instances serve as the basis for the training set. The distribution of tweet instances comes from all English disaster-related tweets events in Crisislex merge to create a single big dataset. These tweets already cleaned up means from noises such as punctuations, URL's, etc.

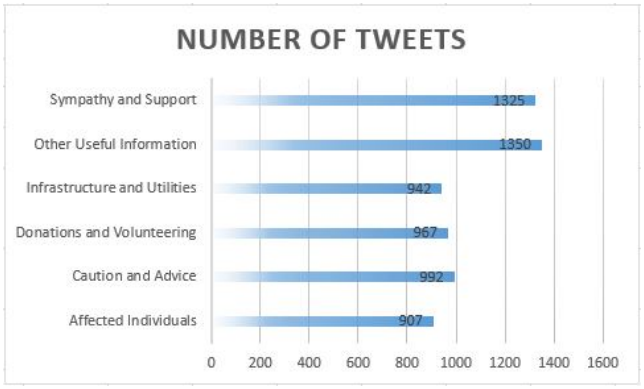
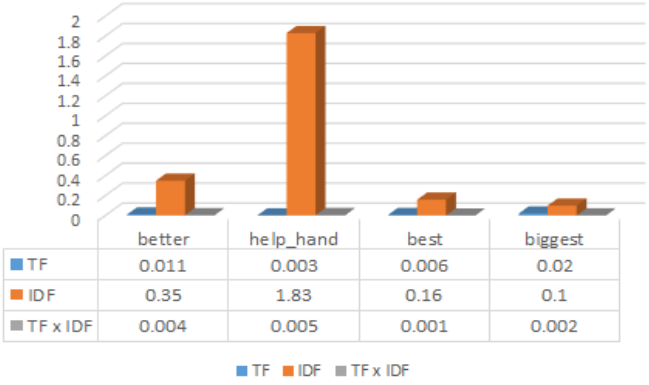


Figure. 4: Number of Tweets Instances Per Class

Figure 5 shows the graph with an example outcome of the TF-IDF collocation process. The diagram shows the sample terms with their resultant TF-IDF weights. As reflected in the example, collocation help_hand has the top TF-IDF score, which means that this term is a proper candidate term to characterize the document. The TF-IDF weight is calculated based on the equation illustrated in figure 3. The TF-IDF for the term "help_hand" is calculated based on the number of appearances of specific terms / total number of terms in a document. As back to the example, 1/350 multiplied by IDF, which is a logarithm of a total number of documents / total number of documents containing the specific term, which is log(2225/33), which gives a TFIDF score of 0.005.



	better	help_hand	best	biggest
TF	0.011	0.003	0.006	0.02
IDF	0.35	1.83	0.16	0.1
TF x IDF	0.004	0.005	0.001	0.002

Figure 5: TF-IDF Collocation Matrix

Figure 6 displays the result of the precision evaluation metric, which explains that there are a visible increase in precision on the three (3) algorithms as applied to two TF-IDF approaches. The precision increase varies from 6% to 23%. The result is obtained by subtracting the precision result of modified TF-IDF less traditional TF-IDF. The traditional TF-IDF serves as the baseline result since traditional TF-IDF performs the single term weight computation. The graph further discloses that RandomForest has a consistent increase in precision by 12% higher than the two other compared algorithms.

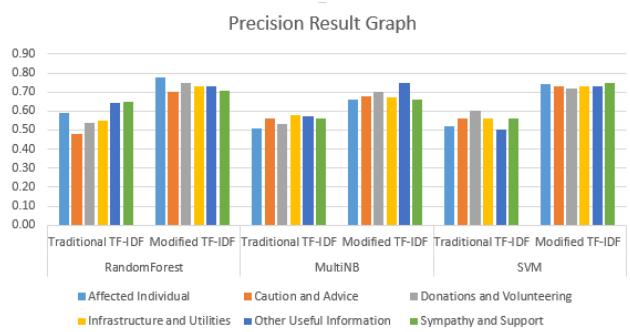


Figure 6: Comparison of Precision Result in Both Traditional TF-IDF and Modified TF-IDF

Figure 7 describes the result of the recall evaluation metric, which clarifies that there is a noticeable escalation in recall on the three (3) considered algorithms, which varies from 4% to 24% as subjected to both traditional TF-IDF and modified TF-IDF process. The graph is further worth noting RandomForest has a constant increase in the recall by 15% higher as compared to other considered classifiers.

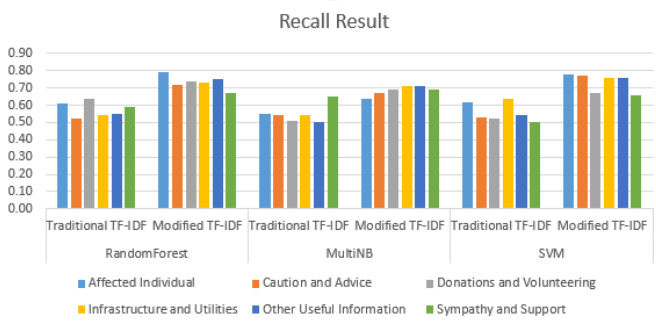


Figure 7: Comparison of Recall Result in Both Traditional TF-IDF and Modified TF-IDF

The graph, as labeled as figure 8, demonstrates the product of the F1 score assessment metric, which exposes that the three (3) considered classifiers, as fed to two TF-IDF approaches, has an apparent increase in F1 score ranging from 7% to 22%. The diagram further discusses that RandomForest has consistently outperformed the two other classifiers by 15% higher.

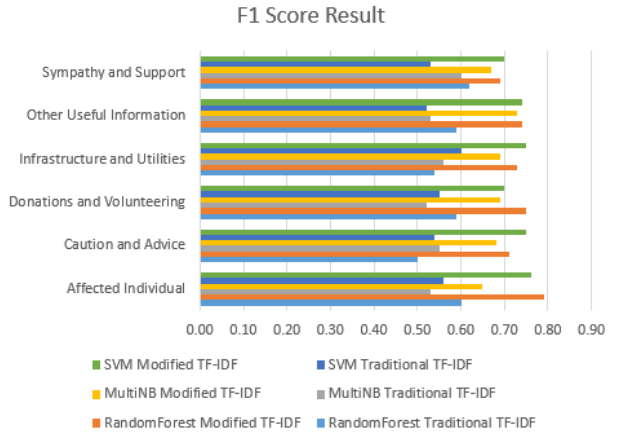


Figure 8: Comparison of F1 Score Result in Both Traditional TF-IDF and Modified TF-IDF.

Figure 9 displays the feature terms with the highest TF-IDF values on one (1) of the sample tweet, in which these include three (3) collocated terms, namely super_typhoon, major_earthquake, death_toll. The lowest collocated term value among the top collocated term is the super_typhoon, with a TF-IDF value of 0.59. The rest of the collocated term weight is above 0.73. The diagram further emphasizes that single terms still have the most segment in the entire collection of weighted terms since collocated terms always fewer occurrences as compared to individual terms. The number one (1) highest TF-IDF value is the evacuation with a TF-IDF score of 0.87, which means that this term is rarely seen in the
entire collection of tweets.

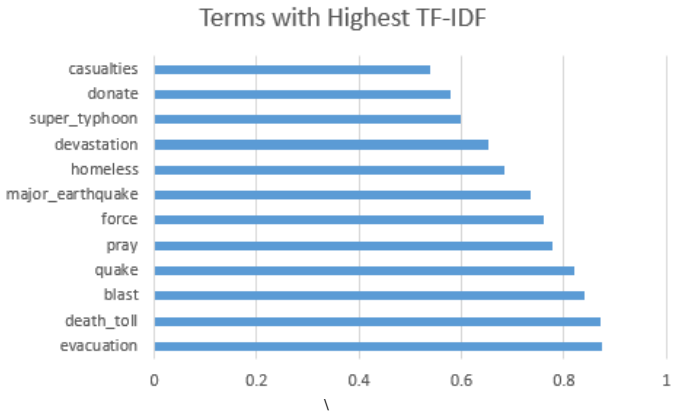


Figure 9: Top Terms with Highest TF-IDF value

5. CONCLUSION

The TF-IDF plays a crucial part in picking the right term that can best describe the document. Hence, selecting the appropriate terms has a significant effect on improving classification effectiveness. This study has fruitfully validated that integrating collocation in the TF-IDF as part of enhancement has supported to increase the classification effectiveness, as presented in the text classifiers evaluation as applied to two TF-IDF approaches, the traditional and modified TF-IDF. The consideration of feature size length text also plays a part in reducing non-informative terms and only retaining the contributing terms.

The study can further enhance classification effectiveness by considering a well-balanced dataset, name entity recognition (NER) inclusion, topic modeling, and dimensional reduction technique.

REFERENCES

1. B. E. Parilla-Ferrer, P. L. F. Jr, and J. T. B. Iv, "**Automatic Classification of Disaster-Related Tweets,**" in International Conference on Innovative Engineering Technologies (ICIET'2014) Dec. 28-29, 2014, Bangkok (Thailand), 2014, pp. 62–69.
2. M. Nguyen, A. Kitamoto, and T. Nguyen, **Advances in Knowledge Discovery and Data Mining**, vol. 9078, no. August. Cham: Springer International Publishing, 2015.
3. R. A. Catanghal Jr, T. D. Palaoag, and A. R. Malicdem, "**Crowdsourcing Approach For Disaster Response Assessment**," MATTER Int. J. Sci. Technol., vol. 3, no. 1, pp. 59–69, May 2017. https://doi.org/10.20319/Mijst.2017.31.5969
4. J. R. Ancheta, C. Sy, L. Maceda, N. Oco, and R. Roxas, "**Computer-assisted thematic analysis of Typhoon Fung-Wong tweets,**" in IEEE Region 10 Annual International Conference, Proceedings/TENCON, 2017, vol. 2017-Decem, pp. 2014–2017.
5. K. Stowe, M. J. Paul, M. Palmer, L. Palen, and K. Anderson, "**Identifying and Categorizing Disaster-Related Tweets,**" in Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media, 2016, pp. 1–6.
6. C. H. Chen, "**Improved TFIDF in big news retrieval: An empirical study,**" Pattern Recognit. Lett., vol. 93, pp. 113–122, 2017. https://doi.org/10.1016/j.patrec.2016.11.004
7. Man Lan, Sam-Yuan Sung, Hwee-Boon Low, and Chew-Lim Tan, "**A comparative e study on term weighting schemes for text categorization,**" in Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., 2016, vol. 2, no. September, pp. 546–551.
8. W. Wang and Y. Tang, "**Improvement and Application of TF-IDF Algorithm in Text Orientation Analysis,**" in Proceedings of the 2016 International Conference on Advanced Materials Science and Environmental Engineering, 2016, no. Amsee, pp. 230–233.
9. S. Siddiqi and A. Sharan, "**Keyword and Keyphrase Extraction Techniques: A Literature Review,**" Int. J. Comput. Appl., vol. 109, no. 2, pp. 18–23, Jan. 2015.
10. W. Zhang, T. Yoshida, and X. Tang, "**A comparative study of TF*IDF, LSI, and multi-words for text classification,**" Expert Syst. Appl., vol. 38, no. 3, pp. 2758–2765, Mar. 2011.
11. F. Olayah and W. Alromima, "**Automatic machine learning techniques (AMLT) for Arabic text**

classification based on term collocations," J. Theor. Appl. Inf. Technol., vol. 96, no. 12, pp. 3729–3738, 2018.
12. J. Violos, K. Tserpes, I. Varlamis, and T. Varvarigou, **Text Classification Using the N-Gram Graph Representation Model Over High-Frequency Data Streams,**" Front. Appl. Math. Stat., vol. 4, Sep. 2018.
13. P. A. Savyanavar and B. Mehta, "**Multi-Document Summarization Using TF-IDF Algorithm.,**" Int. J. Eng. Comput. Sci., vol. 5, no. 4, pp. 16253–16256, Apr. 2016. https://doi.org/10.18535/Ijecs/v5i4.40
14. K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. E. Barnes, and D. E. Brown, "**Text Classification Algorithms: A Survey,**" Information, vol. 10, no. 4, p. 150, Apr. 2019.
15. M. Arshi Saloot, N. Idris, L. Shuib, R. Gopal Raj, and A. Aw, "**Toward Tweets Normalization Using Maximum Entropy,**" in Proceedings of the Workshop on Noisy User-generated Text, 2015, pp. 19–27.
16. [16] M. Shirakawa, T. Hara, and S. Nishio, "**N-gram IDF,**" in Proceedings of the 24th International Conference on World Wide Web - WWW '15, 2015, pp. 960–970. https://doi.org/10.1145/2736277.2741628
17. L. Michelbacher, "**Multi-Word Tokenization for NaturalLanguage Processing,**" Fak. Univ. Inform. Elektrotechnik und Informationstechnik, vol. Ph.D., pp. 1–190, 2013.
18. H. To, S. Agrawal, S. H. Kim, and C. Shahabi, "**On Identifying Disaster-Related Tweets: Matching-Based or Learning-Based?,**" in 2017 IEEE Third International Conference on Multimedia Big Data (BigMM), 2017, pp. 330–337.
19. H. M and S. M.N, "**A Review on Evaluation Metrics for Data Classification Evaluations,**" Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 2, pp. 01–11, Mar. 2015. https://doi.org/10.5121/ijdkp.2015.5201
20. E. Garcia, "**Cosine Similarity Tutorial,**" Inf. Retr. Intell., pp. 4–10, 2015.
21. G. Dalaorao, A. Sison, R. Medina, "**Integrating Collocation as TF-IDF Enhancement to Improve Classification Accuracy,**" in Proceedings - 2019 IEEE The 13th International Conference On Telecommunication Systems, Services, And Application, 2019. https://doi.org/10.1109/TSSA48701.2019.8985458