Section A-Research paper
ISSN 2063-5346

A Proposal of Framework Improving Sentiment Classifier using TF-IDF for the Twitter Dataset and the Tweet Length

[1] Hoong-Cheng Soong, [2] Ramesh Kumar Ayyasamy, [3] Nur Syadhila Che Lah
[1] Department of Digital Economy Technology (DDET), [2][3] Department of Information Systems (DIS),
[1][2][3] Faculty of Information and Communication Technology (FICT), Universiti Tunku Abdul Rahman (UTAR),
Kampar, Malaysia.
[1] soonghc@utar.edu.my, [2] rameshkumar@utar.edu.my, [3] syadhila@utar.edu.my

Abstract— Sentiment Analysis and Opinion Mining are relevant nowadays in many sectors to determine sentiment polarity towards an entity or its aspects. It provides a high percentage of the forecast of the triumph of something, be it events, products, organisations, persons and many more, especially the opinions retrieved from the eminent social media such as Facebook, Twitter or others. Nonetheless, many of the researchers focus on the techniques themselves in the classifying methods without discussing more of the pre-processing parts for the improvement to improve the accuracy based on the corpus sizes and the tweet length or even on the word-embedding or text vectorisation before the passing the tasks to the sentiment classifiers using the myriad of machine learning/deep learning methods. Pre-processing methods, particularly in Natural Language Processing (NLP) stages such as stopwords removal, lemmatisation, stemming, and others, should be done because it is related to the Sentiment Analysis. TF-IDF stands for "Term Frequency — Inverse Document Frequency." and is essential in text retrieval methods to emphasize the crucial words with different weightage in the documents that are undoubtedly helpful to Sentiment Analysis during Word Embedding or Text Vectorisation stages. Furthermore, the tweet length and data sizes as the corpus should identify the effects on the accuracy of the Sentiment Analysis. It is noted that currently, Twitter has allowed from 140 characters to 280 characters for the tweet length, which is interesting to discuss for the Sentiment Analysis using the Twitter dataset. In short, this research proposed several options if the corpus has fewer or fewer datasets and, on the contrary, with massive datasets.

Index Terms— ***Sentiment Analysis and Opinion Mining, Feature Extraction, TF-IDF, Twitter Dataset.***

I. INTRODUCTION

Sentiment Analysis and Opinion Mining are gaining polarity exponentially in this era due to the globalization of information and technology, which is a data-driven to operate crucially in business sectors [21]. Therefore, astute digital entrepreneurs will often take precautions to consider the public opinions or sentiments towards their products or services to survive in an arduous business ecosystem [19]. Thus, this field has gained immense popularity among the ardent researchers to divulge numerous approaches, a variety of novel frameworks and even impressive techniques. Not only beneficial in the business sector, but sentiment analysis or opinion mining is also practically advantageous to other sectors such as politics, health departments, educational institutions, entertainment industries and others [19].

Before differing further, numerous recent researchers often consider Sentiment Analysis and Opinion Mining are parallel to each other. Hence, we comprehend that although in this field, the terms mentioned earlier are used interchangeably [20], both perform different things although closely associated with each other. In short, opinion mining is initially accomplished by discerning the phrases or words consisting of people's opinions that have the sentiment that exclude the objective statements that are considered as facts that are not useful to the sentiment analysis [20]. Meanwhile, views (opinions) held by the opinion holders furnish the sentiments for sentiment analysis, regarded as subjectivity analysis [20]. Succeed the opinion mining steps [20], and Sentiment Analysis is supplementary then provided to the analysts whether the extensive group of people have the feeling to mutually like (positive), to showing dislike (negative) or probably to stay neutral towards an entity or aspect such as product, service, people, organization or many others. For instance, let's take an apparent product Samsung Galaxy S22 Ultra, as an example of the concise explanations of the battery life, screen, weight, cover, and sound quality as the aspects of the aforementioned entity. John expressed, "I adore using Samsung Galaxy S22 because it is trendy. However, the battery life is short. Samsung's latest phone supports 5G". In this context, John is the opinion holder and has positive sentiment toward the word 'adore' on the phone. However, it is unfortunate to have a negative feeling about the aspect of the entity, which is battery life is 'short'. It is considered an objective statement (not applicable to Sentiment Analysis) to the state supporting 5G, which will be excluded from the sentiment analysis.

Across the vast journal databases [14], you can discover numerous techniques or methodologies proposed or researched by diverse renowned researchers that consist of predominantly two sections that is lexicon approaches which are practically using statistical models to determine the sentiments from the corpus (dictionary/dataset) and the other one using computational methods that are none other than machine learning/deep learning approaches under the Artificial Intelligence domain. Apart from that, Sentiment Analysis apparently diverged into four approaches document, sentence, phrase and aspect-based, as the most detailed and focused as distinctly compared to the document-based to generally analyse the entire document [14]. As in figure 1, it is essential to have a bird-eye view for the sentiment analysis taxonomy to further comprehend the sentiment analysis and opinion mining discussed in this paper [16][17][18]. Most researchers focus on the machine learning/deep learning methods due to the disadvantages of using lexicon-based methods with tedious and laborious tasks to classify the sentiments [23]. There are customarily supervised-learning, or unsupervised learning methods within machine learning/deep learning methods, which both have advantages that supervised learning is more accurate and reliable [24]. Still, unsupervised-learning methods outperform supervised-learning methods in terms of less complexity and real-time analysis [24].

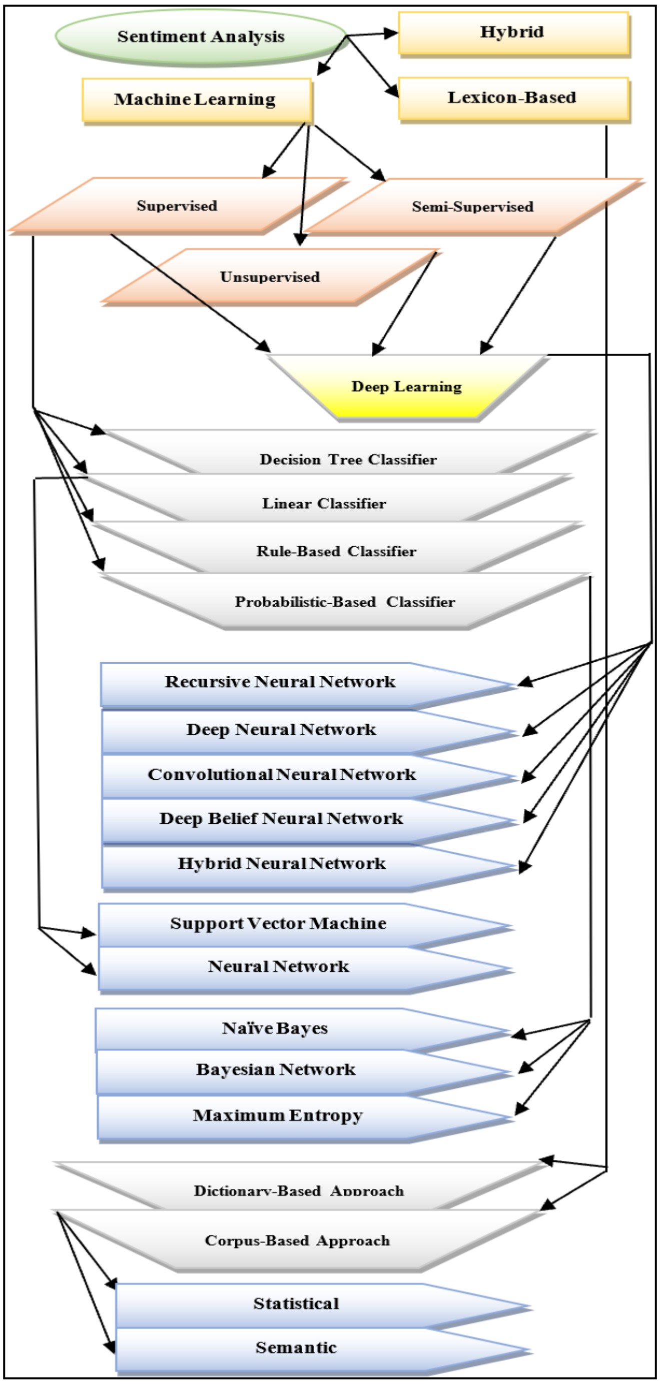


Fig. 1. Sentiment Analysis Taxonomy. Source: [16][17][18]

Please be noted that throughout the entire discussion, focus on text-only because the models, techniques or framework discussed might not be pertinent to image, audio and video. Multimodal approaches [25], either photo or audio or hybrid, require disparate methods or techniques due to the different nature of multimedia elements compared to texts. Before immersing in the in-depth discussion for the subsequent chapters, please be noted that Sentiment Analysis or Opinion Mining patently consists of two significant segments. It performs Natural Language Processing (NLP) at the initial stage and thereon classifies using machine learning/deep learning from the non-lexicon approaches [16]. Accordingly, the algorithms for deep learning methods can perform feature extractions automatically as if employing deep learning without feature extractions [26]. In addition, deep learning sentiment analysis is getting popular that either can be using Recurrent Neural Networks (RNN) or Convolutional Neural Network (CNN) methods. An RNN is trained to distinguish patterns across time, while a CNN studies to identify patterns across space [28]. RNN and CNN exhibits the different architectures that CNN is using to resolve the spatial data for images meanwhile RNN is analyzing sequential or temporal data that is more appropriate for text even for video [28]. However, both methods are capable to classify the polarities for the sentiment analysis and many of the researchers combined the methods (hybrid) [22][27] to give the best possible results in terms of speed, accuracy or precision.

II. Related Works

As Since the pandemic of COVID-19 since initially reported to the World Health Organization (WHO) on December 31, 2019, there has been a massive growth in the topic using the sentiment analysis among the researchers can be obtained from the journal databases. For instance, Marcec and Likic [1] used the lexicon-based method (statistical model) to perform the sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. Apart from that, AlBadani, Shi and Dong [2] proposed a novel machine learning method to perform the sentiment analysis incorporating the universal language model fine-tuning and Support Vector Machines (SVM). For the deep learning method, Singh et al. [3] suggested the sentiment analysis of COVID-19 reviews using the proposed Long short-term memory- Recurrent Neural Networks (LSTM-RNN)-based network. Since the combination of deep learning algorithms are getting relevant nowadays, Zhang and Wu [4] combined the CNN methods with the LSTM which it is the extension from the RNN method for the semantic analysis. Since not all the words share the same weightage and importance, TF-IDF is essential to improve the sentiment analysis. You may scrutinize the Patil's and Kohe's research work [5]. As for Chiny et al. [6], they proposed myriad of approaches combined from lexicon based, non-lexicon based, machine learning and deep learning sentiment analysis as in LSTM, Vader and TF-IDF based hybrid sentiment analysis model. Another latest concerning TF-IDF with Sentiment Analysis by Liu et al. can be found in this paper [50]. Nonetheless, you may examine of our previous survey or review papers [7][8] concerning fundamentals of sentiment analysis or opinion mining for a better insight into this field. In addition, you may browse through the latest review papers concerning the sentiment analysis and opinion mining [9][10][11][12][13][14][15].

III. Proposed General Frameworks

Refer to Figure 2, it is just a proposed generic framework for performing sentiment analysis with explanations or justifications in each of the sub-sections below.

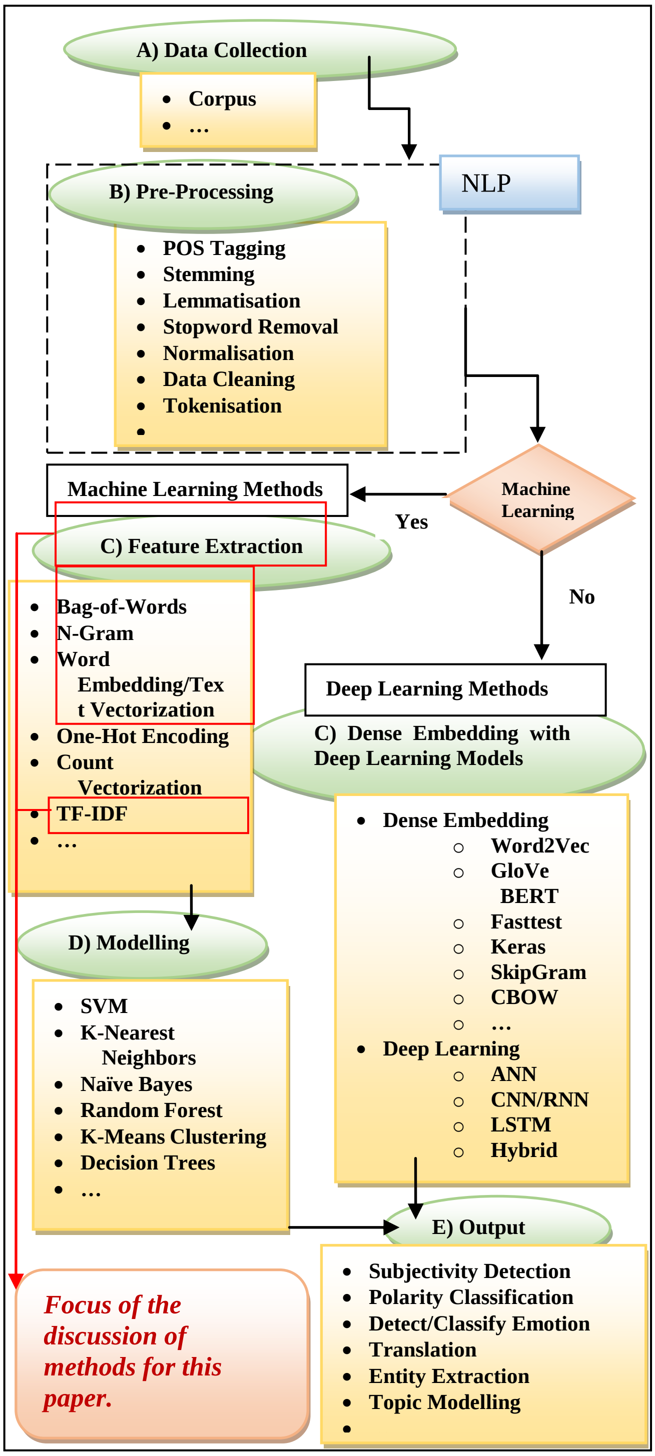


Fig. 2. Proposed Generic Sentiment Analysis Framework.

Before diverging into two sections of the possible methods, either machine learning or deep learning methods, this research paper emphasizes the focal part of the feature extraction (TD-IDF) in the subsequent section, as in figure 2, highlighted with red boxes and arrows. Due to the limited time, the label 'dense embedding with deep learning model' will be our future work to propose another dense word embedding method. The future proposed work will be used compare relatively with the current proposed TF-IDF methods in this research paper for the evaluations to examine meticulously which one will perform better. Sometimes, there are various or newer methods can be found and therefore the list is not final and exhaustive to list the possible methods in figure 2.

A. Data Collection

Heretofore, the data set for the sentiment analysis is typically received from the corpus. A corpus is a collection of digital texts obtained from a database [31]. The particular interest in the data set for my research will be the Twitter dataset because

Twitter posts are commonly brief and constantly generated by the public that is excellently adapted for sentiment analysis and opinion mining [30].

B. Pre-Processing

During the pre-processing stages, the data must be cleansed and standardized to remove symbols, emojis, short forms and a mixture of foreign languages or others to provide better data set to be fed into the machine learning or deep learning algorithms for the classifications. Whenever there are case-sensitive issues [32], it is most reasonable to transform the text into lower cases for standardisation, and sometimes normalisation of words is overlooked; for instance, "goooood" or "gud" share the same meaning as "good". In the normalization phase [32], Stemming and lemmatisation under normalisation offer the benefit of reducing the number of tokens in the model. Both share the same task of minimising the suffixes in the root word. For example, transform "loving", "loves", "loved" to "love". However, stemming is less precise because the rule-based approach removes the suffix directly. For instance, "studies" to become "studi" might misinterpret the word. Lemmatisation is dictionary-based when the meaning of the words will be much more accurate, translating "studies" to root from to "study" accurately [32]. Stopword removal will remove the typical set of words used in English that is not significant in the analysis, such as "is", "the", "for", "in" and many more [32]. Meanwhile, tokenisation separates the texts into smaller units related to machine learning fundamentals for count vectorisation and a transformer for deep learning [32]. It can be derived from sentences, characters or subwords. For example, "I love Samsung" can be tokenised as three words ["I", "love", "Samsung"]. If needed, Part-Of-Speech (POS) tagging is used to tag and to identify the nouns, verbs, adverbs, adjectives and others in the sentences [32].

C. Feature Extraction or Dense Embedding with Deep Learning Models

When using machine learning approaches, feature extraction is crucial for classifying the sentiments in the modelling stage with the preferred algorithms [33]. On the other hand, the deep learning model will automatically perform the feature extractions in the dense layer of the network in a more straightforward meaning it is not a requirement [26]. Although text vectorisation and word embedding share the same goal, they are utilised to bridge the human and computer or human together to comprehend the input from the user into the machine [34]. Text vectorisations usually focus on translating texts into numerical representations for the computer [34]. In contrast, word embeddings concentrate on extracting the valuable features from the texts to be represented by numbers in the form of vectors [34]. The proposed TF-IDF methods clarify the steps in detail in the next section to avoid repetitive explanations in this section. As you notice, the feature extractions and modelling are placed in the same stages as mentioned before for the deep learning methods. The feature extractions are automatically performed in the dense layer network such as CNN, RNN, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) and many more. Even hybrid methods are possible [27].

D. Modelling

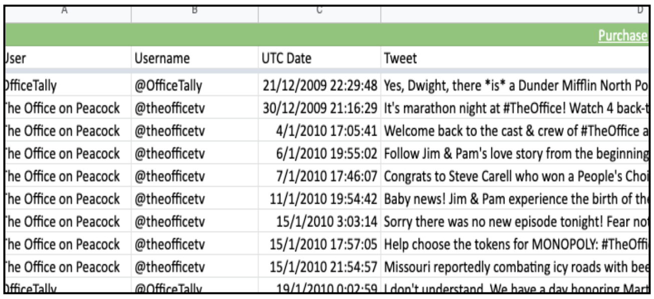
Sentiment classification is the computerised process of determining opinions in a text and thus tagging them as positive, negative, or neutral [35]. Usually, a rule-based model, such as a statistical or semantic model, is straightforward, but the tasks are laborious and time-consuming. Hence, numerous machine learning methods such as Naive Bayes, Random Forest, K-Nearest Neighbors, SVM, Decision Trees and many others are used to automate the process and even possible hybrid methods to increase the effectiveness [35].

E. Output

The output provides many features for the users, not limited to sentiment classification. The users employ the sentiment analysis system for entity detections, emotion detections, topic modelling, spam detection, subjectivity detections, extractions, and many others [36]. In short, the recommended sentiment analysis evaluations examine the accuracy, precision, recall and F1 scores apart from visualising the sentiment scores or rating from the system. Methodologies [37].

IV. PROPOSED FEATURE EXTRACTION METHODS (TF-IDF)

As in figure 3, the Twitter tweets occupy each row in the document, usually in the excel file as one document in a row with additional details such as hashtags, users, timestamps and others [38]. Most of the historical Twitter dataset is available as long it remains as public [38].



A	B	C	D
			Purchase
ser	Username	UTC Date	Tweet
fficeTally	@OfficeTally	21/12/2009 22:29:48	Yes, Dwight, there *is* a Dunder Mifflin North Po
he Office on Peacock	@theofficetv	30/12/2009 21:16:29	It's marathon night at #TheOffice! Watch 4 back-t
he Office on Peacock	@theofficetv	4/1/2010 17:05:41	Welcome back to the cast & crew of #TheOffice a
he Office on Peacock	@theofficetv	6/1/2010 19:55:02	Follow Jim & Pam's love story from the beginning
he Office on Peacock	@theofficetv	7/1/2010 17:46:07	Congrats to Steve Carell who won a People's Choi
he Office on Peacock	@theofficetv	11/1/2010 19:54:42	Baby news! Jim & Pam experience the birth of the
he Office on Peacock	@theofficetv	15/1/2010 3:03:14	Sorry there was no new episode tonight! Fear no
he Office on Peacock	@theofficetv	15/1/2010 17:57:05	Help choose the tokens for MONOPOLY: #TheOffi
he Office on Peacock	@theofficetv	15/1/2010 21:54:57	Missouri reportedly combating icy roads with bee
fficeTally	@OfficeTally	19/1/2010 0:02:59	I don't understand. We have a day honoring Mart

Fig. 3. Sample of Twitter Dataset. Source: [38]

There are no definite rules that you have to set the length of the tweets to perform analysis. Nevertheless, it affects the sentiment analysis outcome if the tweets' lengths are too short, leaving out most of the details to capture the users' sentiments. In addition, the author [39] stated that the longer the tweets, the higher the mean of the sentiments, and it implies better analysis results if the tweets are longer and more detailed. Initially, Twitter limited the length from 140 characters and now increased it to 280 characters currently as the form of microblogging to ensure the users are on point without long-winded unnecessary points. The doubled increment to 280 characters currently can be inferred to avoid short forms, symbols, emojis or jargon or even a mixture of foreign languages to describe their opinions. Selecting only tweets' lengths of 100 and above is suggested to ensure non-bias sentiment is captured on one side from the short tweets' length unintentionally on a similar topic [39]. The shorter the length, the fewer opinions will be captured.

Let's examine two phrases to apprehend the feature extraction for sentiment analysis. The phrase "The Samsung Galaxy has the best camera." and the other phrase "The iPhone 14 Pro has the best display." In sentiment analysis, "Samsung Galaxy" and "iPhone 14 Pro" will be extracted as entities, "camera" and "display" will be the aspects and the word "best" is a positive sentiment. Word Embedding or Text Vectorization is divided into two distinct categories [40]: 1) Frequency-Based for machine learning, such as Bag-Of-Words (BOW), TF-IDF, Count Vector and others. 2) Prediction-Based for deep learning such as Word2Vec, BERT, SkipGram, CBOW (Continuous Bag of Words) and others. It is encouraged to look at this research paper [40] to understand further the aforementioned methods. Before executing the vectorisation, the two phrases mentioned earlier will be tokenized as terms first, as shown in Table 1 and the usual way is to represent the word in binary using the concept of BOW instead of raw counts weighting variant that is similar to term frequency (TF) and count vector before feeding into the machine learning algorithms. BOW does not consider the order of the terms.

TABLE I. BOW FOR THE VECTORSATION

Terms	**Phrases**	
	Document, D1	***Document, D2***
The	1	1
Samsung Galaxy	1	0
iPhone 14 Pro	0	1
has	1	1
best	1	1
camera	1	0
display	0	1

The concept of N-Gram will be handy in this situation, as shown in Table 1. The size for the recommended N-Gram to predict the occurrence of words will be trigram [41] represented in (1) because if it uses a unigram, it is less meaningful if you separate "Samsung" and "Galaxy" and same goes to “iPhone”, “14” and “Pro”.

$$P(w_{1,n}) = P(w_n|w_{n-2}, w_{n-1}) \qquad (1)$$

Please note that it uses different TF weighting scheme for the results in table 2 that under the category of Boolean or binary weighting scheme variant. Nonetheless, the formula shown in (2) is the usual term frequency variant. TF-IDF is better to be employed on top of BOW because some common words, such as the stopword "the", frequently found in the document, contribute less significantly to the sentiment analysis. TF-IDF is a numerical statistic that distinguishes how a word is vital to the documents. The TF-IDF equations are represented in (2), (3), (4).

$$TF = \frac{Frequency\ of\ Words\ in\ Document}{Total\ Number\ of\ Words\ in\ Document} \qquad (2)$$

$$IDF = Log(\frac{Total\ Documents}{Document\ Contain\ the\ Word}) \qquad (3)$$

$$TF - IDF = TF * IDF \qquad (4)$$

With the representation of equations (2), (3) and (4), the vectorization for TF-IDF is shown in table 2. The value TF-IDF is higher it implies the term is more relevant and when it has the value 0 it means it is not relevant [41].

TABLE II. TF-IDF FOR THE VECTORSATION

Terms	**TF**		***IDF***	**TF - IDF**	
	Document, D1	***Document, D1***		***Document, D2***	***Document, D2***
The	1	1	Log(2/2)	0	0
Samsung Galaxy	1	0	Log(2/1)	0.301	0
iPhone 14 Pro	0	1	Log(2/1)	0	0.301
has	1	1	Log(2/2)	0	0
best	1	1	Log(2/2)	0	0
camera	1	0	Log(2/1)	0.301	0
display	0	1	Log(2/1)	0	0.301

As mentioned in an earlier paragraph concerning the tweet lengths, Twitter sentiment analysis deals with fewer and small data sets that will cause the same effect as a "cold start" in recommender systems that are analogous to sentiment analysis [42][49]. On the other hand, there are indeed limitations for machine learning and deep learning to deal with an enormous Twitter dataset size. Recommendations to deal with the large Twitter datasets in sentiment analysis:

- Real-Time sentiment analysis that collects the new Twitter dataset from time to time [43].
- Perform batch processing for the sentiment analysis instead of process in one shot [44].
- Divide and conquer to split the into smaller fragmented datasets and combined the results at later stages [45].
- Big Data applications for the sentiment analysis that can deal with the unstructured and semi-structured data [46].

The output of the sentiment analysis shall be tested and evaluated by using accuracy, precision, recall and F1 score [37][48] as to find out the performance of the Sentiment Analysis for the improvements as represented in the formulas as in (5), (6), (7) , (8) and (9). The general method to calculate the accuracy for sentiment analysis for the correct rating is:

[Num. of Correct Queries / Total Num. of Queries] (5)

Precision has the formula:

P (Pos) = TP/(TP+FP), P(Neg) = TN/(TN+FN) (6)

F1 has the formula:

2*((precision*recall)/(precision+recall)) (7)

Accuracy:

(TP+TN)/(TP+TN+FP+FN) (8)

Recall:

R(Pos) = TP/(TP+FN), R(Neg) = TN/(TN+FP) (9)

Where, T=True, P=Positive, F=False and N=Negative.

CONCLUSION

Based on the results of the sample calculation in table 1 and table 2, not all words share the same weightage of importance and relevance before feeding into machine learning or deep learning algorithm for the sentiment classifications. Furthermore, pre-processing for the NLP tasks is also essential for both machine learning and deep learning for better results after cleaning the data during pre-processing [32]. Some researchers argued that smaller data sets work well with higher accuracy using machine learning instead of deep learning sentiment analysis [43]. The advantage of performing sentiment analysis using the deep learning method is that feature extractions are performed automatically in the dense layer of the network [26]. Please note that feature extraction can enhance the accuracy of the learning algorithm and shrink the time required for the analysis [47]. Since the output of the sentiment analysis provides many types of visualisations such as entity detections, polarity classifications, topic modelling and others, sentiment analysis evaluations, namely accuracy, precision, recall and F1 score, are crucial to measuring the effectiveness of the sentiment analysis system [37][48].

ACKNOWLEDGMENTS

I would like to thank my supervisor, Dr. Ramesh Kumar Ayyasamy and co-supervisor, Dr. Nur Syadhila Che Lah, Faculty of Information and Communications Technology (FICT) in Universiti Tunku Abdul Rahman (UTAR) for the support with the UTARRF grant, UTARRF/2020-C2/S10.

REFERENCES

[1] R. Marcec and R. Likic, “Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna Covid-19 Vaccines,” *Postgraduate Medical Journal*, vol. 98, no. 1161, pp. 544–550, 2021.

[2] B. AlBadani, R. Shi, and J. Dong, “A novel machine learning approach for sentiment analysis on Twitter incorporating the universal language model fine-tuning and SVM,” *Applied System Innovation*, vol. 5, no. 1, p. 13, 2022.

[3] C. Singh, T. Imam, S. Wibowo, and S. Grandhi, “A deep learning approach for sentiment analysis of COVID-19 reviews,” *Applied Sciences*, vol. 12, no. 8, p. 3709, 2022.

[4] W. Zhang and Y. Wu, “Semantic sentiment analysis based on a combination of CNN and LSTM model,” *2022 International Conference on Machine Learning and Knowledge Engineering (MLKE)*, 2022.

[5] R. S. Patil and S. R. Kolhe, “Supervised classifiers with TF-IDF features for sentiment analysis of Marathi tweets,” *Social Network Analysis and Mining*, vol. 12, no. 1, 2022.

[6] M. Chiny, M. Chihab, O. Bencharef, and Y. Chihab, “LSTM, Vader and TF-IDF based hybrid sentiment analysis model,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, 2021.

[7] H.-C. Soong, N. B. Jalil, R. Kumar Ayyasamy, and R. Akbar, “The essential of sentiment analysis and opinion mining in social media : Introduction and survey of the recent approaches and Techniques,” *2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, 2019.

[8] H.-C. Soong, R. K. Ayyasamy, and R. Akbar, “A review towards deep learning for sentiment analysis,” *2021 International Conference on Computer & Information Sciences (ICCOINS)*, 2021.

[9] S. Verma, “Sentiment analysis of Public Services for Smart Society: Literature Review and Future Research Directions,” *Government Information Quarterly*, vol. 39, no. 3, p. 101708, 2022.

[10] R. Kaur and S. Kautish, “Multimodal sentiment analysis,” *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*, pp. 1846–1870, 2022.

[11] S. T. Al-Otaibi and A. A. Al-Rasheed, “A review and comparative analysis of sentiment analysis techniques,” *Informatica*, vol. 46, no. 6, 2022.

[12] G. D’Aniello, M. Gaeta, and I. La Rocca, “KnowMIS-ABSA: An overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis,” *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5543–5574, 2022.

[13] L. Zhu, M. Xu, Y. Bao, Y. Xu, and X. Kong, “Deep learning for aspect-based sentiment analysis: A Review,” *PeerJ Computer Science*, vol. 8, 2022.

[14] M. Wankhade, A. C. Rao, and C. Kulkarni, “A survey on sentiment analysis methods, applications, and challenges,” *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022.

[15] A. Nazir, Y. Rao, L. Wu, and L. Sun, “Issues and challenges of aspect-based sentiment analysis: A comprehensive survey,” *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 845–863, 2022.

[16] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, “Sentiment analysis based on Deep Learning: A Comparative Study,” *Electronics*, vol. 9, no. 3, p. 483, 2020.

[17] B. K. Bhavitha, A. P. Rodrigues, and N. N. Chiplunkar, “Comparative study of machine learning techniques in sentimental analysis,” *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2017.

[18] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[19] F. Rapport, R. Clay-Williams, and J. Braithwaite, *Key concepts in implementation science: Translation and improvement in medicine and Healthcare*. Abingdon, Oxon: Routledge, 2022.

[20] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.

[21] M. A. Ullah, K. Munmun, F. Z. Tamanna, and M. S. Chowdhury, “Sentiment analysis using ensemble technique on textual and Emoticon Data,” *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, 2022.

[22] K. Pasupa and T. Seneewong Na Ayutthaya, “Hybrid deep learning models for Thai sentiment analysis,” *Cognitive Computation*, vol. 14, no. 1, pp. 167–193, 2021.

[23] K. Aggarwal, “Has the future started? the current growth of artificial intelligence, Machine Learning, and Deep Learning,” *Iraqi Journal for Computer Science and Mathematics*, pp. 115–123, 2022.

[24] P. Monika, C. Kulkarni, N. Harish Kumar, S. Shruthi, and V. Vani, “Machine learning approaches for sentiment analysis,” *International journal of health sciences*, pp. 1286–1300, 2022.

[25] X. Yan, H. Xue, S. Jiang, and Z. Liu, “Multimodal sentiment analysis using multi-tensor fusion network with Cross-modal modeling,” *Applied Artificial Intelligence*, vol. 36, no. 1, 2021.

[26] Y.-C. Huang, T.-H. Chuang, and C.-J. Lin, “A revisit for the diagnosis of the hollow ball screw conditions based classification using Deep Learning,” *Measurement and Control*, 2022.

[27] W. Ukaihongsar and W. Jitsakul, “Enhancing sentiment analysis using hybrid deep learning,” *Proceedings of the 18th International Conference on Computing and Information Technology (IC2IT 2022)*, pp. 183–193, 2022.

[28] V. Karas and B. W. Schuller, “Deep learning for sentiment analysis,” *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*, pp. 27–62, 2022.

[29] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning and Data Mining*. Boston: Springer, 2017.

[30] A. Quazi and M. K. Srivastava, “Twitter sentiment analysis using machine learning,” in *Lecture Notes in Electrical Engineering*, Singapore: Springer Nature Singapore, 2023, pp. 379–389.

[31] S. Uday Sampreeth Chebolu, F. Dernoncourt, N. Lipka, and T. Solorio, *Survey of Aspect-based Sentiment Analysis Datasets. arXiv e-prints*. 2022.

[32] S. M. Emaduddin, R. Ullah, I. Mazahir, and M. Z. Uddin, "Enhancing Information Preservation in Social Media Text Analytics Using Advanced and Robust Pre-processing Techniques," *International Journal of Media and Information Literacy*, vol. 7, no. 1, pp. 60–70, 2022.

[33] S. Styawati, A. Nurkholis, A. A. Aldino, S. Samsugi, E. Suryati, and R. P. Cahyono, "Sentiment analysis on online transportation reviews using Word2Vec text embedding model feature extraction and support vector machine (SVM) algorithm," in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, 2022.

[34] H. D. Abubakar, Dept. of Computer Science, Jigawa State Colledge of Education, Gumel, Nigeria, M. Umar, and Dept. of Computer Sceince, Faculty of Science, Sokoto State University, Sokoto, Nigeria, "Sentiment classification: Review of text vectorization methods: Bag of words, tf-idf, Word2vec and Doc2vec," *SLU Journal of Science and Technology*, vol. 4, no. 1 & 2, pp. 27–33, 2022.

[35] K. Gulati, S. Saravana Kumar, R. Sarath Kumar Boddu, K. Sarvakar, D. Kumar Sharma, and M. Z. M. Nomani, "Comparative analysis of machine learning-based classification models using sentiment classification of tweets related to COVID-19 pandemic," *Mater. Today*, vol. 51, pp. 38–41, 2022.

[36] M. Hao *et al.*, "Visual sentiment analysis on twitter data streams," in *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2011.

[37] R. R. Subramanian, N. Akshith, G. N. Murthy, M. Vikas, S. Amara and K. Balaji, "A Survey on Sentiment Analysis," *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 70-75, 2021.

[38] "Twitter historical data," *Tweet Binder*, 21-Dec-2021. [Online]. Available: https://www.tweetbinder.com/blog/twitter-historical-data/. [Accessed: 15-Sep-2022]

[39] M. Mayo, "A clustering analysis of tweet length and its relation to sentiment," *arXiv [cs.CL]*, 2014.

[40] M. A. H. Wadud, M. F. Mridha, and M. M. Rahman, "Word Embedding methods for word representation in deep learning for Natural Language Processing," *Iraqi J. Sci.*, pp. 1349–1361, 2022.

[41] P. Gupta, S. Nigam and R. Singh, "A Ranking based Language Model for Automatic Extractive Text Summarization," *2022 First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR)*, pp. 1-5, 2022.

[42] F. G. Contratres, S. N. Alves-Souza, L. V. L. Filgueiras, and L. S. DeSouza, "Sentiment analysis of social network data for cold-start relief in recommender systems," in *Advances in Intelligent Systems and Computing*, Cham: Springer International Publishing, pp. 122–132, 2018.

[43] A. P. Rodrigues *et al.*, "Real-time Twitter spam detection and sentiment analysis using machine learning and deep learning techniques," *Comput. Intell. Neurosci.*, 2022.

[44] P. Harnmetta and T. Samanchuen, "Sentiment analysis of Thai stock reviews using transformer models," in *2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2022.

[45] W. Yu Chung Wang, D. Pauleen, and N. Taskin, "Enterprise systems, emerging technologies, and the data-driven knowledge organisation," *Knowl. manag. res. pract.*, vol. 20, no. 1, pp. 1–13, 2022.

[46] W. G. Mutasher and A. F. Aljuboori, "Real time big data sentiment analysis and classification of Facebook," *Webology*, vol. 19, no. 1, pp. 1112–1127, 2022.

[47] M. Suhaidi, R. Abdul Kadir, and S. Tiun, "A review of feature extraction methods on machine learning," *Journal of Information System and Technology Management*, vol. 6, no. 22, pp. 51–59, 2021.

[48] A. Saxena, H. Reddy, and P. Saxena, "Introduction to Sentiment Analysis Covering Basics, Tools, Evaluation Metrics, Challenges, and Applications," in *Principles of Social Networking. Smart Innovation, Systems and Technologies*, vol. 246, A. Biswas, R. Patgiri, and B. Biswas, Eds. Singapore: Springer, 2022.

[49] B. J. Goh, H.-C. Soong, and R. K. Ayyasamy, "User song preferences using Artificial Intelligence," *2021 IEEE International Conference on Computing (ICOCO)*, 2021.

[50] H. Liu, X. Chen, and X. Liu, "A study of the application of weight distributing method combining sentiment dictionary and TF-IDF for text sentiment analysis," *IEEE Access*, vol. 10, pp. 32280–32289, 2022.

AUTHORS PROFILE

Mr Hoong-Cheng Soong is a postgraduate student and at the same time as a lecturer in Universiti Tunku Abdul Rahman (UTAR) in Malaysia. Currently, he is currently pursuing PhD in UTAR under the tutelage of Dr Ramesh Kumar Ayyasamy and Dr Nur Syadhila Che Lah in the Sentiment Analysis field and his background initially is majoring in Database Management. He received his Bachelor's and Master's degree in Computer Science from Universiti Teknikal Malaysia Melaka (UTeM) in Malaysia. Currently, his research interests are not limited to Sentiment Analysis from the Natural Language Processing field but also in Computational Music, Graphical Passwords Security, Machine Learning or Deep Learning and Audio Processing. He is a student member of IEEE and a certified Professional Technologist by Malaysia Board of Technologists (MBOT).

Dr Ramesh Kumar Ayyasamy (SMIEEE) received the Ph.D. degree in Information Technology from Monash University, Australia, in 2013. Starting from 2003-2008, and 2013-2014 he worked as a lecturer in TamiI Nadu, India, and in Monash University, Malaysia consecutively. Since 2015, he is working as an Assistant Professor in Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Malaysia. His research interests include Artificial intelligence, Big Data Analytics, Cyberbullying, Deep Learning, Machine Learning, and Text Mining. Dr. Ramesh is senior member of IEEEE, and a member of the International Association of Computer Science and Information Technology (IACSIT).

Dr Nur Syadhila Che Lah was born in Penang, Malaysia, on November 8, 1987. She received the B.S. degree in Computer Sciences, in 2009 and the M.S. degree in Information Technology Management, in 2013 from Universiti Teknologi Malaysia (UTM). She obtained a Ph.D. in Information Systems from Universiti Teknologi Malaysia (UTM) in 2020. She is currently working as an assistant professor in the Faculty of Information and Communication Technology (FICT), Universiti Tunku Abdul Rahman (UTAR), Perak Campus. Her research interest includes Information Retrieval, Persuasive Systems and E-Commerce.