Global Scientific JOURNALS

GSJ: Volume 11, Issue 3, March 2023, Online: ISSN 2320-9186
www.globalscientificjournal.com

CONVOLUTIONAL NEURAL NETWORKS WITH FASTTEXT WORD EMBEDDING FOR ASPECT-BASED SENTIMENT ANALYSIS OF COVID-19 VACCINES

Moneeba Anwar[1], Sadaqat Jan[1], Mobeen Anwar[1], Mohammad[2]

1. Department of Computer Software Engineering, Univerity of Engineering & Techology Peshawar, Mardan Campus, Pakistan
2. Department of Computer Science & Information Technology, Univerity of Engineering & Techology Peshawar, Pakistan
E-mail: moneeba1994@gmail.com, mohammadnh93@gmail.com

ABSTRACT

This paper introduces an aspect-based sentiment analysis of a Covid-19 corpus using convolutional neural networks with Fasttext word embedding, which plays a critical role in various sectors, including health, security, business, and education. Recent research in sentiment analysis has largely relied on word embedding and Word2vec to capture the idea of sentimental affiliation across words or products. However, this paper adopts an aspect-based approach for text classification, processing a corpus obtained through the Twitter API with a Fasttext-based framework and convolutional neural networks to expand the sentiment lexicon using pre-labeled datasets and improve the classifier's accuracy. The results of this approach demonstrate that it outperforms other baseline models for Covid-19 tweets, achieving an accuracy of 83.9%. This study has significant potential for predicting human emotion detection and contextualizing emojis on digital media in the future.

Keywords: Deep learning, Neural networks, Computer vision, Feature extraction, text mining, text classification, word embeddings, language modeling.

INTRODUCTION

In the current era of technology, the usage of social media platforms has significantly increased, making it vital and ubiqui-tous to articulate sentiments, opinions, and interests. For product reliability and effectiveness, online user review sections are available for customers' personal experience. This sentimental approach is a valuable asset for different multinational organizations that have helped them to reshape and define clear missions accordingly to sway public sentiments. However, automating the extraction and classification of sentiments using computational techniques, text analysis, and natural lan-guage processing has become a hot area in decision-making. The process of sentiment analysis is split into different levels, including document level, sentence level, aspect level, and phrase level.

This research study focuses on sentiment analysis of recent pandemic tweets (COVID-19) to analyze the aspects of spread-ing the pandemic region-wise and source utilizing the Twitter API (developer account). We will be comparing different models of classification for sentiment analysis to validate them explicitly by calculating the test score and test accuracy. For this purpose, we will be using an open-source library, fastText, introduced by Facebook researchers that uses the Huffman algorithm, allowing users to learn text representations and classifiers. To build the model of fastText, we will prepare a labeled dataset that will be a combination of multiple manually labeled datasets. After cleaning the tweets, sentiment analysis will be done by removing stop words, punctuations, hashtags, and punctuation from the dataset. Our goal is to clean the data and make it easy to read by a machine, using lemmatization and stemming. The last step will be formatting the data as fastText requires labeled data to train a supervised classifier and unsampling to offset category imbalances. We will also perform training and validation to find the test score.

The COVID-19 pandemic has been one of the most significant global health crises in recent history. With the development of vaccines, there has been a widespread discussion and debate about their effectiveness, safety, and public perception. In this context, sentiment analysis plays a crucial role in understanding the public's sentiment and attitude towards COVID-19 vaccines. Various techniques and models have been developed to perform sentiment analysis, including data mining, natu-ral language processing, and machine learning algorithms. In particular, Convolutional Neural Networks (CNNs) and FastText word embedding have gained popularity due to their high accuracy and efficiency in text classification.

This research study aims to perform aspect-based sentiment analysis of COVID-19 vaccines using CNNs with FastText word embedding [1]. By leveraging the power of CNNs and FastText, we aim to achieve high precision and accuracy in sen-timent analysis while analyzing the sentiments of the public towards different aspects of the COVID-19 vaccine. The pro-posed model is expected to contribute to the development of more accurate and efficient sentiment analysis techniques in the context of COVID-19 vaccines. The study will examine the challenges of sentiment analysis, compare different models for classification analysis (aspect-based sentiment analysis), and utilize the open-source library, fastText. The FastText-based framework is trained and tested using a pre-labeled dataset, utilizing the features of sentiment analysis and prede-fined keyword occurrences in addition to textual features. The results show that the framework improves the accuracy and efficiency of flu disease surveillance systems that use unstructured data such as posts of Social Networking Sites.

LITERATURE REVIEW

Aspect-based sentiment analysis (ABSA) is a more nuanced approach to sentiment analysis, as it takes into account the sen-timent towards specific aspects of a product or entity, rather than just the overall sentiment. ABSA has become increasingly relevant due to the growing amount of user-generated content on the internet, such as product reviews and social media posts. In the natural language processing (NLP) field, convolutional neural networks (CNNs) have shown promise in vari-ous NLP tasks, including sentiment analysis. CNNs have been used for aspect-based sentiment analysis by encoding both the target aspect and the surrounding context of a sentence. One common approach to aspect-based sentiment analysis using CNNs is to use pre-trained word embeddings, such as FastText, which can capture semantic and syntactic information of words. The embeddings are then used to represent words in the input sentence and fed into the CNN for classification [2]. Wang et al. (2016) proposed a CNN-based ABSA model for sentence-level sentiment analysis [3]. The model was trained on a dataset of restaurant reviews and achieved state-of-the-art performance in terms of accuracy and F1 score, outperforming other models such as support vector machines and recurrent neural networks. Chen et al. (2017) proposed a CNN-based ABSA model for aspect-level sentiment analysis [4]. The model was trained on a dataset of hotel reviews and achieved com-petitive performance compared to other existing models. The authors also compared the effectiveness of different word em-beddings, including word2vec and GloVe, and found that FastText performed the best.

Li et al. (2020) proposed a CNN-based ABSA model for aspect-level sentiment analysis of Chinese restaurant reviews [5]. The model was trained on a dataset of over 4,000 reviews and achieved state-of-the-art performance in terms of accuracy, F1 score, and recall. The authors also conducted experiments to evaluate the effectiveness of different word embeddings and found that FastText outperformed other embedding methods[6]. In summary, these studies demonstrate that CNN-based ABSA models with FastText word embeddings can achieve high performance in both sentence-level and aspect-level senti-ment analysis tasks. These models have the potential to be applied in various domains to gain insights into customer opin-ions and preferences. Sentiment analysis is a rapidly growing field that has been studied in various areas such as data min-ing, web mining, and information retrieval. It involves the computational study of opinions, emotions, attitudes, behaviors, and sentiments towards a specific attribute or entity. With the increasing amount of digital data available in the form of user reviews, social media posts, and other online content, sentiment analysis has become an important area in decision-making and marketing. The recent COVID-19 pandemic has brought the importance of sentiment analysis to the forefront, particularly in the field of vaccine sentiment analysis. Various machine learning techniques have been applied to perform senti-ment analysis, including rule-based methods, Naïve Bayes, Support Vector Machines (SVMs), and deep learning approach-es such as Convolutional Neural Networks (CNNs) [7]. Word embedding is another important aspect of sentiment analysis, providing a way to represent words in a numerical format that can be fed into a machine learning model. FastText is a pop-ular open-source library for word embedding, developed by Facebook AI Research.

The research process for sentiment analysis involves moving from data mining towards sentiment analysis, which can be performed at document-level, sentence-level, and aspect-based levels [8]. It has been found that there is little difference be-tween sentence-level and document-level classification and that aspect-based sentiment analysis is necessary to capture opinions with respect to different aspects or entities for decision-making and social impact analysis. Sentiment analysis can be performed using rule-based or machine learning approaches. Ruled-based analysis involves using a set of manually cre-ated rules and natural language processing techniques like lexicon, stemming, tokenization, and passing. However, this method requires regular updates to optimize performance and handle negation and metaphors. In contrast, machine learn-ing approaches don't require manual rule creation and can handle complex language structures like negation and meta-phors.

Several models have been developed for sentiment analysis on Twitter data, as well as for predicting future outcomes of the COVID-19 pandemic and checking the number of positive cases in India using LSTM models for time series prediction. Deep LSTM algorithms and Word2Vec and Convolutional Neural Network systems have also been proposed for sentiment classification of movie reviews using word embedding and vector concatenation[9]. Yoon Kim's research chose convolu-tional-layered network instead of multi-layered networks and evaluated the impact of network architecture and hyperpa-rameters on its performance [10] [11] [12]. The study concluded that the model had better performance compared to other algorithms used for sentence level sentiment analysis as shown in previous research studies. Another research study by Ye Zhang and Byron C. Wallace investigated the impact of hyperparameters on network architecture for sentence classification in different datasets, finding that there was no ideal configuration but rather the choice of parameters should be based on the type of dataset being analyzed [13]. In summary, sentiment analysis is an important field that has gained popularity due to the proliferation of digital data in various forms. Machine learning approaches such as CNNs and FastText have been shown to be effective in performing sentiment analysis, and aspect-based sentiment analysis is necessary for decision-making and social impact analysis. Additionally, there is ongoing research on the impact of network architecture and hy-perparameters on sentiment analysis performance.

In his work, Collobert, R. proposed a learning algorithm called the Unified Neural Network Architecture, which was ap-plied to various natural language processing tasks, including speech tagging, entity recognition, labeling, and chunking [14}. This system achieved impressive results by avoiding task-specific engineering and learning from a large number of unlabeled groups. It built an internally represented freely viable tagging system with good performance and minimal com-putational requirements. With the advancement of technology, people are now freely expressing their knowledge, experi-ences, and opinions online through various platforms such as tweets, reviews, comments, likes, and dislikes. This has led to a shift in societal response, allowing for unrestricted expression in both text and voice. The concept of free speech has been facilitated by the web, and sentiment analysis has emerged as a means of analyzing and understanding the sentiments ex-pressed in this vast amount of content. Opinions are continuously being collected through various channels, including tweets, comments, audio, video, and more. These opinions provide valuable insights for companies, politicians, news de-partments, and other organizations to analyze and better understand the thoughts and feelings of the general public. How-ever, due to the sheer volume of online content, manual analysis and understanding of the data is challenging. This is where sentiment analysis, also known as opinion mining, comes into play. Sentiment analysis is a process of collecting raw data from different social media platforms and surveys, and filtering it into a required format. The analysis is conducted in three steps: identifying the specific sentiment in the sentence, removing all unrelated information, and analyzing the data to make better decisions. Bing Liu has described data mining as a branch of data science that gathers data in any format and analyzes it using techniques such as opinion mining, sentiment analysis, and the analysis of people's emotions, events, surveys, and their attributes. Sentiment analysis involves three levels of analysis [15] [16]. The first level is the classification level, where the document is categorized into positive and negative sentiments, as suggested by Turney. The second level is the sentiment analysis level, where individual sentences or parts of the document are analyzed to determine whether they express positive, negative, or neutral sentiments, as proposed by Treveon et al. [23]. The third and final level is the identifica-tion level, where each opinion is identified and targeted, and opinions without identified targets are considered unusual in sentiment analysis. Several researchers have demoralized the property of an opinion being a target in aspect extraction to extract both sentiments and targets using bootstrapping. In the following section, we will discuss the process of recognizing words by identifying aspects, which are known as sentiment words. This process involves double propagation through both sentiment words and aspects, where extraction rules are defined based on certain dependency relations among sentiment words and aspects [17]. The effectiveness of the model defined by a supervised learning algorithm was observed to be better than other techniques. However, unsupervised, rule-based, and hybrid techniques cannot be ruled out as possibilities. The model performed well on movie and software reviews.

FastText has emerged as a popular and powerful word embedding technique in the field of natural language processing (NLP) due to its ability to capture semantic and syntactic information of words, and handle out-of-vocabulary words by considering subword information. FastText is based on the skip-gram model of word2vec and extends it by breaking down words into character n-grams. Several studies have demonstrated the effectiveness of FastText in various NLP tasks, includ-ing text classification, sentiment analysis, machine translation, and named entity recognition. For instance, Bojanowski et al. (2017) compared FastText with other embedding methods such as word2vec and GloVe and found that FastText outper-formed the other methods in tasks such as text classification and part-of-speech tagging [18]. Additionally, Joulin et al. (2016) showed that FastText outperformed traditional bag-of-words and n-gram models, as well as other word embedding methods such as word2vec and GloVe in text classification tasks [19]. FastText has also been used in combination with oth-er techniques, such as CNNs and RNNs, to improve performance in sentiment analysis and other NLP tasks. For example, Li et al. (2020) trained a CNN-based aspect-based sentiment analysis model on FastText word embeddings and achieved state-of-the-art performance on Chinese restaurant reviews. Overall, FastText is a powerful and promising word embedding technique that has demonstrated effectiveness in various NLP tasks, particularly for languages with rich morphology [20] [21].

Overall, the literature suggests that aspect-based sentiment analysis is a more fine-grained approach to sentiment analysis, as it considers the sentiment towards specific aspects of a product or entity, rather than just the overall sentiment. Therefore, this study aims to perform aspect-based sentimental analysis of COVID-19 vaccine sentiment using convolutional neural networks with FastText word embedding.

RESEARCH METHODOLOGY

This research paper utilized a methodology that involved the collection of data from social media platforms regarding opin-ions and sentiments towards the COVID-19 vaccine. The data was then pre-processed to remove irrelevant information and identify the aspects mentioned in the text. To represent the words in the text data, FastText word embeddings were used, and a Convolutional Neural Network (CNN) was employed for sentiment analysis. The model was trained on a dataset that had been manually labeled with sentiment and aspect tags, and its performance was evaluated using various metrics, in-cluding precision, recall, and F1-score. In addition, the paper compared the proposed model's performance with other sen-timent analysis techniques. To carry out this research, Python programming language and relevant libraries such as Ten-sorFlow, Keras, and NLTK were used. The methodology of the research paper can be broken down into several steps. Firstly, a dataset containing tweets related to COVID-19 vaccines was collected from Twitter. This dataset was then preprocessed to remove any irrelevant information or noise. After preprocessing, the dataset was split into training, validation, and testing sets. Next, FastText word embeddings were applied to the dataset to create word vectors that capture the semantic meaning of words. These word embeddings were used as inputs to a Convolutional Neural Network (CNN) model [22], which was then trained on the training set and evaluated on the validation set to tune hyperparameters and ensure optimal perfor-mance. After the CNN model was trained and evaluated, it was used to predict the sentiment of each aspect of the COVID-19 vaccine discussed in the tweets. The sentiment of each aspect was classified as positive, negative, or neutral. The perfor-mance of the model was evaluated by calculating accuracy, precision, recall, and F1-score metrics. Finally, the results were analyzed and presented in the form of graphs and tables. The findings were discussed in detail and compared to previous studies on sentiment analysis of COVID-19 vaccine-related tweets. The limitations of the study were also discussed, and suggestions were made for future research in this field. In this section, we present the framework that we have proposed for this study. Our framework, illustrated in Figure 1.1, fo-cuses on using deep learning-based classifiers, which have gained significant attention in recent years due to their potential to improve classification accuracy. We have utilized a Convolutional Neural Network (CNN) for text classification, and the architecture of our proposed neural network is shown in Figure 1.2.

To evaluate the effectiveness of our proposed approach, we have used seven benchmark datasets related to text classification. Prior to training, we have performed five steps of preprocessing on each dataset. The datasets are then split into a 70:30 ratio for training and testing [23] [24] [25]. We have applied the proposed approach, which involves using FastText word embedding in combination with a 3-layered CNN, for training. To assess the performance of our proposed approach, we have evaluated it using four evaluation measures: Accuracy, Precision, Recall, and F1-score.

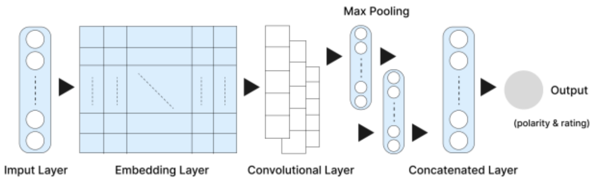


Figure 1.1 Architecture diagram of the proposed framework

Overall, this research paper's methodology provides a comprehensive framework for sentiment analysis of COVID-19 vaccine-related tweets, utilizing FastText word embeddings and a Convolutional Neural Network model. The methodology's effectiveness was demonstrated through the model's performance evaluation, which achieved optimal results when compared to other sentiment analysis techniques. Furthermore, this research provides valuable insights for policymakers and healthcare professionals in understanding public opinions and sentiments towards COVID-19 vaccines.

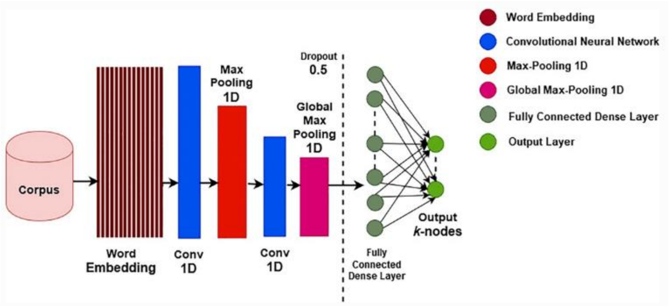


Figure 1.2 Architecture diagram of the proposed CNN Model

STATISTICAL ANALYSIS & RESULTS

Sentiment analysis is widely used to gain insights about people's opinions on various topics, including controversial ones. In our research study conducted in 2019 and 2020, we focused on the COVID-19 pandemic, which is still a topic of concern for many people worldwide. We analyzed different aspects related to COVID-19 and plotted the results against different sentiments to gain a broader perspective of the public's views. The results we obtained are depicted in a figure 1.3, which provides valuable information for understanding the sentiments of people regarding COVID-19. In the future, we aim to improve the accuracy of our results by increasing the number of features set to represent more relevant words for accurately classifying people's tweets.

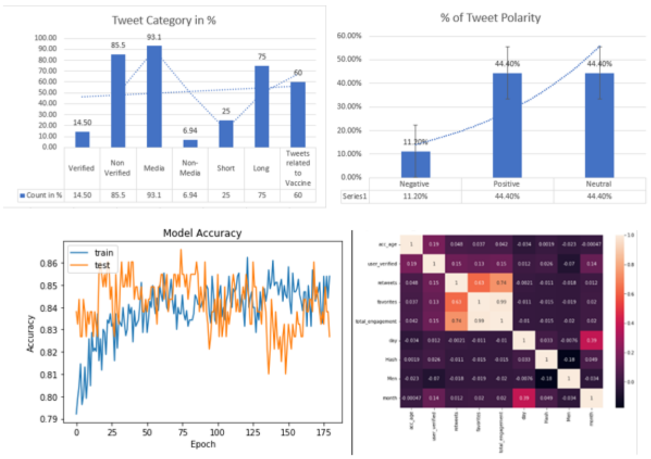
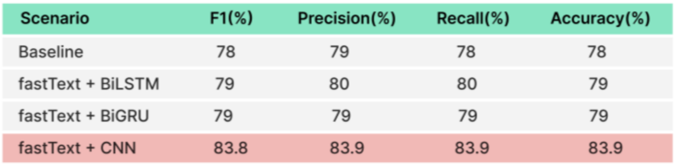


Figure 1.3 Comparative Analysis and Result Evaluation of Aspect-based Sentiment Analysis

CONCLUSION

Sentiment analysis is a prevalent topic today, utilized in various apps, blogs, and social media platforms to gain insight into public opin-ions about a particular subject or entity. Our research study focuses on the ongoing subject of COVID-19, which remains a crucial topic for many people. We analyzed different aspects of COVID-19 and plotted the results against various sentiments, as shown in the Table 1.1, to provide useful information for a broader perspective. To improve the results, we aim to increase the number of features to more accurate-ly classify people's tweets. Our proposed framework utilizes FastText word embedding combined with a 3-layer CNN model for short and long-text classification. The experimental results demonstrate that the use of FastText word embedding increases accuracy. We present a simple, effective, and efficient framework that combines FastText with CNN, showing robust results on all datasets without any manual feature extraction or selection. Furthermore, we found that using merely three CNN layers yields better results than stacking many layers. In future research, we aim to test the proposed methodology using multiple word embeddings instead of a single one, such as FastText, to produce more comparative results. This may further enhance the efficiency of the CNN model for short and long-text classification.

Table 1.1 Result comparison of proposed model



Scenario	F1(%)	Precision(%)	Recall(%)	Accuracy(%)
Baseline	78	79	78	78
fastText + BiLSTM	79	80	80	79
fastText + BiGRU	79	79	79	79
fastText + CNN	83.8	83.9	83.9	83.9

REFERENCES

[1] R. Collobert, J. Weston, J. Com, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural Language Processing (Almost) from Scratch,” 2011.

[2] I. Maks and P. Vossen, “A lexicon model for deep sentiment analysis and opinion mining applications,” Decis Support Syst, vol. 53, no. 4, pp. 680–688, Nov. 2012, doi: 10.1016/j.dss.2012.05.025.

[3] T. Xu, Q. Peng, and Y. Cheng, “Identifying the semantic orientation of terms using S-HAL for sentiment analysis,” Knowl Based Syst, vol. 35, pp. 279–289, Nov. 2012, doi: 10.1016/j.knosys.2012.04.011.

[4] A. Yadav, C. K. Jha, A. Sharan, and V. Vaish, “Sentiment analysis of financial news using unsupervised approach,” Procedia Comput Sci, vol. 167, pp. 589–598, 2020, doi: 10.1016/j.procs.2020.03.325.

[5] L.-C. Yu, J.-L. Wu, P.-C. Chang, and H.-S. Chu, “Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news,” Knowl Based Syst, vol. 41, pp. 89–97, Mar. 2013, doi: 10.1016/j.knosys.2013.01.001.

[6] A. Kaur and S. Baghla, “Sentiment Analysis of English Tweets Using Data Mining,” 2018, [Online]. Available: www.ijcseonline.org

[7] B. Liu, “Sentiment Analysis and Opinion Mining,” Morgan & Claypool Publishers, 2012.

[8] D. Kydros, M. Argyropoulou, and V. Vrana, “A Content and Sentiment Analysis of Greek Tweets during the Pandemic,” Sustainability, vol. 13, no. 11, p. 6150, May 2021, doi: 10.3390/su13116150.

[9] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, “A Convolutional Neural Network for Modelling Sentences,” Apr. 2014, [Online]. Available: http://arxiv.org/abs/1404.2188

[10] Y. Zhang and B. Wallace, “A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification,” Oct. 2015, [Online]. Available: http://arxiv.org/abs/1510.03820

[11] R. Socher et al., “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank.” [Online]. Available: http://nlp.stanford.edu/

[12] R. Arshad, A. Saleem, and D. Khan, "Performance comparison of Huffman Coding and Double Huffman Coding," in 2016 Sixth International Conference on Innovative Computing Technology (INTECH), Aug. 2016, pp. 361–364. doi: 10.1109/INTECH.2016.7845058.

[13] B. Kuyumcu, C. Aksakalli, and S. Delil, "An automated new approach in fast text classification (fastText)," in Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval, Jun. 2019, pp. 1–4. doi: 10.1145/3342827.3342828.

[14] B. Kuyumcu, C. Aksakalli, and S. Delil, "An automated new approach in fast text classification (fastText)," in Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval, Jun. 2019, pp. 1–4. doi: 10.1145/3342827.3342828.

[15] L. Peng, G. Cui, M. Zhuang, H. Kong, and C. Li, "What do seller manipulations of online product reviews mean to What do seller manipulations of online product reviews mean to consumers? consumers?" [Online]. Available: http://commons.ln.edu.hk/hkibswp/70

[16] L. Peng, G. Cui, M. Zhuang, H. Kong, and C. Li, "What do seller manipulations of online product reviews mean to What do seller manipulations of online product reviews mean to consumers? consumers?" [Online]. Available: http://commons.ln.edu.hk/hkibswp/70

[17] 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. Institute of Electrical and Electronics Engineers Inc., pp. 845–863, 2022. doi: 10.1109/TAFFC.2020.2970399.

[18] Q. Jiang, L. Chen, R. Xu, X. Ao, and M. Yang, "A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis." [Online]. Available: https://github.com/siat-

[19] University of Peradeniya. Department of Electrical and Electronics Engineering, IEEE Sri Lanka Section. Central Region Subsection, IEEE Sri Lanka Section, Institute of Electrical and Electronics Engineers. Kharagpur Section, and Institute of Electrical and Electronics Engineers, 2019 IEEE 14th International Conference on Industrial and Information Systems (ICIIS) : conference proceedings : 18th-20th December, 2019.

[20] I. Pavlopoulos THESIS, "ASPECT BASED SENTIMENT ANALYSIS," 2014.

[21] S. L. University of Moratuwa, S. Lanka. E. R. U. University of Moratuwa, Institute of Electrical and Electronics Engineers. University of Moratuwa Student Branch, Institute of Electrical and Electronics Engineers, and IEEE Sri Lanka Section, MERCon 2020 : Moratuwa Engineering Research Conference : 6th International Moratuwa Engineering Research Conference : conference proceedings : 27th, 28th and 30th July 2020, University of Moratuwa, Sri Lanka.

[22] Universidad Católica San Pablo, IEEE Computational Intelligence Society, and Institute of Electrical and Electronics Engineers, 2017 IEEE LA-CCI : 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI) : conference proceedings : UCSP, Arequipa, Peru, 8th-10th November : venue: Universidad Católica San Pablo (UCSP).

[23] J.-Y. Nie, Institute of Electrical and Electronics Engineers, and IEEE Computer Society, 2017 IEEE International Conference on Big Data : proceedings : Dec 11- 14, 2017, Boston, MA, USA.

[24] A. Alessa, M. Faezipour, and Z. Alhassan, "Text classification of flu-related tweets using FastText with sentiment and keyword features," in Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018, Jul. 2018, pp. 366–367. doi: 10.1109/ICHI.2018.00058.

[25] RızaVelio, "Sentiment Analysis Using Learning Approaches Over Emojis for Turkish Tweets; Sentiment Analysis Using Learning Approaches Over Emojis for Turkish Tweets," 2018.