SENTIMENT ANALYSIS IN SOCIAL MEDIA AND FINANCIAL DATA LITERATURE REVIEW

Bharathidasan Ilango[1], Giridhar Dhanapal[2], Indrajith Shanmugasundaram[3], Jafreen Kazi[4], Sanjana Nitin Tarekar[5], Shreya Krishnarth[6] and Sivaramakrishnan Sridharan[7]

Department of Computer Science, University of Liverpool, Liverpool L69 3BX.

Abstract

In modern world, sentiment analysis stands out as an important study area in demand within Natural Language Processing (NLP). Social media is the biggest contributor to the views of the people. Opinions taken from one the of the commonly utilised platform i.e., twitter is used to examine the attitudes of people from recent coronavirus epidemic. Data gathered through the collection of postings on the pandemic comprise the extent of dataset of the research. This research is not confined just to the existing pandemic but also gives insights drawn from financial data. This research is used to establish that the popularity of the product is strongly impacted by the opinions of the individuals with high audience access on social media. Some of the contemporary concerns are connected to slang terminology, new dialects, grammar, and spelling faults, etc. The current literature research seeks to analyse roughly 24 publications, which cover distinct types of programmes being used for emotional analysis. Our objective is to review multiple Algorithms with several datasets not only from twitter but also from other sources. We try to establish the accuracy of the models reviewed and determine the best model to reach our goal.

Keyword: *Sentiment Analysis; Machine Learning Algorithms; Neural networks; Big Data; Artificial intelligence.*

Abbreviation Table

Abbreviations	**Descriptions**
MaxEnt	Maximum Entropy
SVR	Support Vector Regression
DTs	Decision Trees
BERT	Bidirectional Encoder Representations from Transformers
NLP	Natural Language Processing
RF	Random Forest
NB	Naive Bayes
RNN	Recurrent Neural Networks
MCNN	Memory based Convolutional Neural Networks
TF-IDF	Term Frequency for Inverse Document Frequency
LSTM	Long Short-Term Memory
SVM	Support Vector Machine
KNN	K-Nearest Neighbour
LR	Logistic Regression
NLTK	Natural Language Tool Kit
SSWE	Sentiment Specific Word Embedding

AR	Association Rule Classifier
OM	Opinion Mining
DLMNN	Deep Learning Modified Neural Network
IANFIS	Improved Adaptive Neuro Fuzzy Inference Systems
CBOW	Continuous Bag of Words model

1. Introduction

Polarity analysis of a user’s opinions to conduct sentiment analysis on the dataset has evolved as one of the major purposes of all the data monetization firms. This study has proven to be a game changer in a variety of fields, including retail, e-commerce, and blogging. Widely employed, and the greatest platform to acquire data that expresses users’ perspectives is social media. Not just blogging sites such as Twitter, Instagram, Reddit, Facebook, etc., but also retail portals like Amazon, Flipkart, and eBay are used for opinion mining.

The polarity can be expressed as "positive," "neutral," or "negative." This direction of sentiment analysis encompasses text analytics, computational linguistics, and NLP applications for recognising and categorising the perspectives of the user. In general, the primary purpose of the sentiment analysis is to determine the author's point of view related to the identical circumstance or the whole document's contextual polarity. The perspective could either be the user's judgement or assessment or their emotional state.

This review motivated us to address the following questions -
RQ1) How to conduct opinion mining using big data i.e., twitter data?
RQ2) Does the sentiment of tweets by executives have a deeper influence than general tweets?

Models used to answer the above-mentioned questions are Machine Learning model, Lexicon model, Hybrid/Ensemble model. A detailed explanation is provided in the methodology, which can be reviewed in the **Table 1: Various Methods used in Literature review.**

This Literature review is designed in following manner: Section II contains Literatures reviewed and its Related work methodology design Section III contains various methodologies used in reviewed literature Section IV contains Data Extraction methods Section V contains Applications Section VI contains Conclusions Section VII contains Limitations.

2. Literature Review

A. Related Work

The objective of the Saad and Yang study [14] is to employ machine learning algorithms to do a complete sentiment analysis of tweets dependent on Ordinal Regression. The proposed approach starts by pre-processing tweets and thereafter employing a methodology for feature extraction to produce an effective feature. The score and balancing components are then divided into various categories. The proposed approach employs Decision Trees (DT), Support Vector Machine (SVM), Random Forest (RF), Support Vector Regression (SVR) and Multinomial Logistic Regression (SoftMax) [1] algorithms for sentiment analysis categorisation. The actual execution of this system makes advantage of a publicly accessible Twitter dataset supplied via the NLTK corpus resources. They applied their strategy to training data that contains emoticon-filled tweets, which behaved as noisy labels. They created models utilising Maximum Entropy (MaxEnt), Support Vector Machines (SVM) and Naïve Bayesian classifiers [5]. The characteristics include Bigrams, POS, and Unigrams. They observed that SVM performed substantially better than most of the models and that unigram functioned better as feature.

The author advised implementing the sentiment-specific word embedding (SSWE) model in this research [16]. Without utilising any explicit annotations, they developed three neural network models to learn sentiment-specific word embedding (SSWE) from massive, unsupervised tweets. It was the first piece of research to employ word embedding to detect Twitter sentiment. The emotion of text is usually neglected by the algorithms that are presently employed to generate continuous word representations and study the semantic context of words. This provided a problem for sentiment analysis since, normally, words with identical syntactic context but differing emotion polarity, such as "excellent" and "awful," are mapped to close word vectors. By improving the word embedding learning approach (Collobert et al., 2011) and utilising three neural networks to train SSWE, the author suggests incorporating the sentiment information of sentences to generate continuous representations for words and phrases.

LI Xiaojun and his co-authors [17] presented the C&W-SP model, a text sentiment analysis approach based on representation learning. First, a C&W model-based improved word embedding learning approach that takes emotional information and speech patterns into account is recommended. The NLP&CC's 2013 assessment sets of data are used to evaluate experimental outcomes with different models. The experimental findings reveal that the C&W-SP model, which includes emotion and part of speech information, works efficiently, demonstrating the usefulness of the recommended strategy.

L. David and K. Jon addressed the link prediction issue in 2003, and they devised techniques for link prediction [18] depending on how near or remote sites were located to each other in a system. Machine learning (ML) is becoming more and more popular for different categorization and prediction difficulties. Node-pair similarity metrics, which are unsupervised techniques, constitute the basis of the bulk link in prediction research. The unsupervised techniques may be advanced to supervised binary classification in link prediction unless the links are named when the data set is formed. In this research [2], the author described each tweet by merging all syntactic and semantic information, dubbed "hybrid features." The syntactic information was obtained using the bag of words approach, while the semantic information was derived from a combination of domain-specific (DS) methods and fast-text-based (FT) methods. Secondly, they created a unique multi-channel convolutional neural network (MCNN) that aggregates the various CNNs to collect multi-scale information for improved classification. And at last, they tested the feasibility of both the suggested feature extraction approach and the MCNN model categorising tweets as three sentiment categories (positive, neutral, and negative) using the NepCOV19Tweets dataset, which is the sole available COVID-19 tweets set of data in Nepali. The assessment findings reveal that the suggested hybrid features beat individual feature extraction techniques with the greatest classification accuracy of 69.7%, and the MCNN model exceeds the current approaches with the best classification accuracy of 71.3% during classification.

In 2022, Zihan [19] Tang and Xinyu Li carried out tests on Twitter data using a range of models and algorithms, including both traditional deep learning (BERT) and machine learning (NB, CNB, RF, and SVM) methods. Considering the test findings and a theoretical and practical analysis of the benefits and limitations of the approaches investigated, they picked BERT to extract characteristics from the tweets. The daily average sentiment score was computed to construct the sentiment index. To further assess the public's sentiment trend in the pre- and post-Covid eras, they applied the autocorrelation function (ACF), auto regression (AR), and partial autocorrelation function (PACF) based on the sentiment index. In terms of the US, UK, and the whole globe, the regression findings demonstrate a positive correlation of time trend prior to COVID, an irrelevant correlation during COVID, and a negative correlation post-COVID. The assessments of the variety in people's attitudes about inflation, meanwhile, hint both at cross-national and temporal differences. Using an upgraded word embedding [3] approach and an LSTM network for sentiment analysis, Alsayat and Ahmed developed an ensemble deep learning language model in 2022 [20]. When comparing their model to current standards with varied degrees of complexity, the author obtained greater classification performance than the present state-of-the-art sentiment analysis algorithms. Additionally, they examined the model using a Twitter dataset that was notably tied to a coronavirus (COVID-19) outbreak to determine how effectively it may be utilised to predict sentiment via the examination of tweets. The findings revealed that our ensemble technique for sentiment analysis may be effective.

This study, which assessed the impact of social media posts on the close price forecast of shares using tweets and Reddit posts, was suggested by Anubhav Sarkar et al. in 2022 [21]. The idea was to link sentiments via social media platforms with past stock prices and apply time series techniques to examine their influence on closing prices. They conducted extensive tests and comprehensive analyses to investigate the impact of CEO and public tweets on the closing price, employing a variety of deep learning-based models on diverse datasets. Assigning sentiment polarity was done at a far greater degree of granularity in a paper by Zhang L. et al. [22], which conducted sentiment analysis at the entity level in 2011. Second, they employed a distinct approach for identifying polarity since they had to deal with three unique forms of emotion (positive, neutral, and negative), making it hard to directly apply their methodology. Many of the neutral tweets retrieved are opinionated, as the Lexicon-Based Approach's positive and negative classes have poor recall. They noticed those op-ed tweets consequently before any categorization could be done.

In this method [8,] the researchers looked for techniques that operate at a high level of analysis, such as considering the semantic orientation of single words and context-specific valence shifters, but not at the full linguistic analysis (which analyses argument structure or word senses); however, further research in that area is possible. This article was created using the Semantic Orientation Calculator (SO-CAL) that the authors have been working on for a while. They initially gathered sentiment-bearing words, including adjectives, verbs, nouns, and adverbs, and employed them to establish semantic orientation while accounting for valence shifters (intensifiers, down, negation, and irreal markers). The utility and robustness of this lexicon-based method were established. One criticism levelled at lexicon-based strategies is that they are untrustworthy because dictionaries are produced either automatically or by hand-ranking by humans (Andreevskaia and Bergler 2008). They released the findings of different tests in Section 3, "Validating the Dictionaries," which demonstrate that their dictionaries were trustworthy and robust when compared to values given by individuals and to other existing dictionaries (using the Mechanical Turk interface).

In 2013, Nitish proposed a sentiment analysis model to focus mostly on market response [23]. The study focuses on the mobile phone industry. They were asking for an outline of what determines a device's grade. Twitter, one of the most extensively used social media sites, was employed to obtain data by employing tweets regarding the merchandise. People's views on a certain product could be positive, negative, or mixed. They collected the tweets with NLP to identify whether a tweet was positive, negative, or neutral information. They created an ontology to determine the importance of the "tweet" in the statement. Any common traits and the relationships between them are contained in an ontology, a form of domain model.

For determining whether a user's attitude is positive, neutral, or unfavourable, P. Sasikala *et al*. [24] published a study in 2020 that dealt with opinion mining (OM) or sentiment analysis (SA). It collects every user's opinion, perspective, and attitude regarding the associated product A method is presented for the SA of online product customer reviews using Deep Learning Modified Neural Networks (DLMNN). A method for predicting the future of online items is also proposed, which makes use of the Improved Adaptive Neuro-Fuzzy Inferences System (IANFIS). First, as part of the dataset, the data values are divided into grade-based (GB), content-based (CB), and collaboration-based (CLB) settings. Each setting is then sent to review analysis (RA) employing DLMNN, which returns outcomes as negative, positive, or neutral assessments. For producing predictions, IANFIS classifies and adds a scale factor to the result. The recommended work performed better than the existing techniques in the experimental evaluation.

In 2018, Imane El Aloui *et al* [4], Proposed technique comprises of first generating a dynamic vocabulary of words' polarities based on a chosen collection of hashtags that are linked to a certain subject and then, categorising the tweets into different groups by incorporating new characteristics that significantly improve the degree of polarity of a tweet. To validate our method, we classified the tweets related to the US election of 2016. The outcomes of prototype testing have revealed good accuracy in distinguishing positive and negative categories and their sub-classes.

In this study from 2018, Sohaniger *et al*. [5] examined whether Deep Learning models may be enhanced to boost StockTwits sentiment analysis capacity. To convey stock market sentiments on StockTwits, they employed a range of Neural Network Models, including Long Short-term Memory, doc2vec, and Convolutional Neural Networks. Their results indicated that a Convolutional Neural Network is the most accurate model for estimating authors' sentiment in the StockTwits dataset, and Deep Learning algorithms may be applied effectively for financial sentiment research.

3. Methodology

After reviewing more than twenty publications, all the papers display similar presentations and usage of Machine Learning approaches. Refer to **Figure 1: Types of Algorithms in Machine Learning Used in Sentiment Analysis**, Lexicon-based approaches, ensemble methods, and some other deep learning methodologies are coupled with some of the available NLP packages provided by Google, Microsoft, and IBM. All these recommended and investigated ways were presented in the **Table 1: Various Methods used in Literature review.** After investigating all the prior studies and grasping different approaches, we concluded that the final model would have a higher prediction if a deep ensemble learning model was added [6].

CITATIONS	ALGORITHM/METHODS	MODELS	ACCURACY
[2]	MCNN	Hybrid model	71.30%
[4]	Naïve Bayes/maximum entropy/SVM	Machine Learning model/Lexicon model/Hybrid model	90.21%
[5]	SVM/Naïve Bayes/Decision tree	Machine learning	
[17]	C&W model/CBOW/SSWE/n-gram	Machine learning algorithm	75.79%
[16]	C&W Model, Sentiment-specific word embedding	Lexicon based	84.89%
[14]	SVR/DT/RF/SOFTMAX	Machine learning method	91.81%
[15]	Naive Bayes/SVM/Maximum entropy	machine learning algorithm	80%
[18]	Common neighbours/Jaccard's coefficient	Ensemble method	
[19]	(VADER, TF-IDF and BERT)/Naive Bayes/SVM/CNB/MNB/LR/RF/GBT	Ensemble model	79.99%
[20]	Word embedding, LSTM, NLP libraries from Google, IBM, Microsoft	Machine Learning model/Lexicon model/Hybrid model	
[21]	RNN, GRU, LSTM, AE	Machine learning method	
[22]	SVM classifier/ME/FBS/AFBS/LLS/LMS	Lexicon-based Method and Learning-based Method	85.40%
[8]	SVM/unigrams/bigrams	Lexicon Method	78.74%
[24]	Naïve Bayesian/maximum entropy/SVM	Machine learning	
	Latent Dirichlet's Allocation/DNN/CNN/deep CNN/RBM	Machine learning method	

Table 1: Various Methods used in Literature review.

Machine learning-based techniques depend on prominent ML algorithms such as Naïve Bayesian, RNN, CNN, K-nearest neighbour, Support vector machines (SVM) and Support Vector Regression (SVR), decision trees (supervised and unsupervised approaches), and supervised and unsupervised methods are all examples of supervised and unsupervised methods. In the supervised techniques, the model was trained by using multiple tagged texts. The well-known methods for opinion mining analysis are Naïve Bayes classification [7], support vector machines (SVM) [8], and the maximum entropy principle. Unsupervised techniques are applied solely when it is challenging to comprehend labelled training materials.

Lexicon-based technique uses a dictionary of recognisably sentimental terms and a sentiment lexicon. Dictionary-based techniques and corpus-based techniques are the two major kinds. The first one searches for terms used to convey views in the text, then looks for the sentiment polarity in dictionaries. There are various dictionaries, including SentiWordNet [9], but they can also be constructed by hand. Finding sentimental words with a context-specific direction is what the second one implies. It starts with a series of symbols representing opinion phrases and searches through a large corpus to locate additional opinion words. This strategy employs adjectives and verbs as indications of the text's semantic direction.

Ensemble technique integrates state-of-the-art methodology and uses a deep learning framework. This is mostly built upon FastText word embedding technology [10] as well as an LSTM network [11]. The ensemble system, which merges the categories of the neural network models to give a consolidated categorization, is the final step of the automated sentiment analytical approach. The main objective of ensemble techniques is to combine the findings of numerous models to attain a majority vote by applying a "hard" or "soft" voting system. While "soft" voting examined the probability distribution of each class (i.e., how confidently the model identifies the tweets as "positive," "negative," or "neutral"), "hard" voting focused on the classification (i.e., "positive," "negative," or "neutral").

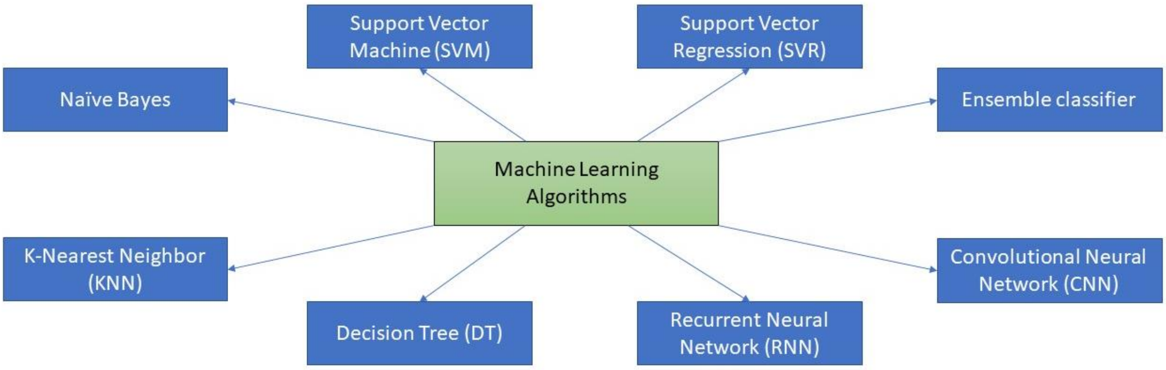
Support Vector Machine (SVM)
Support Vector Regression (SVR)
Naïve Bayes
Ensemble classifier
Machine Learning Algorithms
K-Nearest Neighbor (KNN)
Convolutional Neural Network (CNN)
Decision Tree (DT)
Recurrent Neural Network (RNN)

Figure 1: Types of Algorithms in Machine Learning Used in Sentiment Analysis

The model's accuracy and performance are determined by two independent variables: the total number of hidden neurons and the total number of hidden layers in the network. As data splitting is complete, an ensemble deep learning language is trained with a mix of different hyperparameter values (by employing a method called grid search in data mining). For each set of hyperparameters, multiple models were trained, the accuracy rate on the validation set was examined, and the model with the best accuracy on the testing dataset was chosen. The complete technique is repeated by employing all outstanding folds and datasets. We observe that across all emotion classifiers, our recommended ensemble approach gives greater prediction performance than models in a statistical fashion.

4. Data Extraction

Data mining is a collection of methods for automatically detecting patterns and correlations in massive datasets. One significant benefit of using Twitter data is that it provides numerous APIs that allow developers to use the data. As indicated in [12], Unlike several APIs such as the REST API, Twitter Official API, and Streaming API that cannot access old tweets, PyQuery and snscrape are very few among the Twitter APIs that could even obtain old tweets with much less latency.

Data contributed by users on social media covers a variety of data apart from alphabetic characters such as stop words, usernames, graphical symbols, online connections, punctuation, and URLs. These objects will not contribute to the process of sentiment analysis. For instance, the username doesn't allow any system to adequately distinguish good from bad tweets. Such information was labelled "noise," and its wonderful practise to remove it to boost the effectiveness of classification algorithms. Later in this technique, we delete punctuation, numerals, and undefined characters. In the last step of data analysis, we transform emoticons and visual characters to either negative or positive polarity, then use this translation to assign classifiers to each tweet. As seen in **Table 2: Various Data Sets Used in the Literatures**, after receiving the data, text processing was performed by word segmentation, i.e., the elimination of stop-words using Text Blob [12]. Sentiment analysis evaluates semantic expression and representation. The strength of emotional polarity indicates if the text is strongly disparaging, objective, general derogatory, general praising, or strongly praising.

CITATIONS	DATASET	VOLUME
[14], [23]	Twitter API - mobile product	10,000 tweets
[15]	Twitter emoticon data	1,600,000 tweets
[16]	SemEval 2013 Twitter sentiment classification dataset	12,000 data
[2]	NepCov19Tweets	33,247 total tweets
[19], [20], [21]	Twitter data from 2017 - 2022	140,000 tweets
[22]	Twitter data	972,200 tweets
[24]	Food review dataset	500,000 reviews
[8], [12]	Movie reviews	Text - 800 words
		Movie - 1,900 texts from polarity dataset
		Camera - 2,400 text corpus

Table 2: Various Data Sets Used in the Literatures

We gather the data related to the stock market from Twitter and pre-process it to clean the data. Then tokenization plays a role in identifying the meaning of the sentence or paragraph. Tokenization is a strategy used in natural language processing to divide sentences into smaller parts that can be easily assigned meaning. After pre-processing, the relevant data is retrieved that will aid in analysis.

5. Application

Motivation is to derive opinion mining using financial data to determine if executive's opinions have greater influence over general tweets.

There is a link that shows the usage of executives like the CEOs, celebrities, various social media influencers on Twitter and effect on the stock market activity. In this thesis, author considered multiple hypotheses like – 'The stock market incorporates the effect of tweets of CEOs directly in the stock price.' They considered information and tweets were posted by the most influential top executives CEOs. After the consideration of more than two thousand tweets from the CEO of 500 different companies, it shows the CEO who has larger audience can have direct effect on the rise of the trading volume. These tweets have average positive effect on return. This effect is seen when twitter does not show any lagging traits and the information is perfectly efficient.

6. Conclusion

Our literature review answers the proposed question

RQ1) The proposed ensemble deep learning model shows high potential on sentimental analysis. Whenever a new dataset is presented to this model, it shows a gradual improvement in the performance. So, it can derive the polarity, (i.e., positive, neutral, and negative sentiments) of the dataset with high accuracy.

RQ2) The review suggests that the most influential the executive is, the deeper will be the influence on general people. The impact of the society and the public is created by the executive as compared to the general people, but we cannot consider general people as unimportant. We found that if these models are trained on more granular data with huge real-time dataset, this will outperform the other comparative models [21].

7. Limitations

The data collected from twitter was sufficient to demonstrate the strategy's utility. When compared to circumstances in actual life, the findings were likewise quite accurate. The number of tweets collected was still quite small, however. This occurred because of the location parameter and complex searches.

Social Media Bots (automated accounts) are used to manipulate public opinion and harness the power of social media because of the enormous growth of social media's influence on people's opinions. Academics have developed good algorithms to identify botmasters, but social media bot accounts are constantly changing their bots to avoid detection.

The sample data may be inconsistent, according to this. As a result, underestimating and overestimation become problematic.

CEOs don't post as often as regular individuals; it has been observed. As a result, common tweets and posts were detected in abundance, but executive postings were limited. Furthermore, we solely examined the tweets, including the tickers.

References

[1] B. K. N, S. M. J and N. M, "SoftMax based User Attitude Detection Algorithm for Sentimental Analysis," vol. 125, p. 8, 2018.

[2] S. Chiranjibi and S. B. Tej, "Multi-channel CNN to classify nepali covid-19 related tweets," 2022.

[3] A. A. A and T. Lixin, "Word embeddings for Arabic sentiment analysis," in *IEEE International Conference on Big Data (Big Data)*, Washington, DC, USA, 2016.

[4] A. E. Imane , Y. Gahi, M. Rochdi and C. Youness, "A novel adaptable approach for sentiment analysis on big social data," vol. 5, p. 11, 2018.

[5] D. Wang, K. M. Taghi , A. Pomeranets and S. S, "Big Data: Deep Learning for financial sentiment analysis," vol. 5, no. 1, p. 25, 2018.

[6] A. Ning, D. Huitong, Y. Jiaoyun, A. Rhoda and A. F. Ting, "Deep ensemble learning for Alzheimer's disease classification," vol. 105, p. 11, 2020.

[7] D. Lopamudra, C. Sanjay, B. Anuraag, B. Beepa and T. Sweta, "Sentiment Analysis of Review Datasets using Naïve Bayes' and K-NN Classifier," vol. 8, no. 4, p. 62, 2016.

[8] T. Maite, B. Julian, T. Milan, V. Kimberly and S. Manfred, "Lexicon-Based Methods for Sentiment Analysis," vol. 37, no. 7, p. 42, 2011.

[9] D. Kerstin, "Using SentiWordNet for multilingual sentiment analysis," in *IEEE*, Cancun, Mexico, 2008.

[10] A. Ben, W. G. Andrew and A. Anima, "Probabilistic FastText for Multi-Sense Word Embeddings," vol. 1, no. 1, p. 11, 2018.

[11] C. Nan and W. Peikang, "Advanced Combined LSTM-CNN Model for Twitter Sentiment Analysis," in *IEEE*, Nanjing, China, 2019.

[12] A. Shen, "Sentiment Analysis Based on Financial Tweets and Market Information," in *International Conference on Audio, Language and Image Processing (ICALIP)*, Honkong, China, 2018.

[13] K. Knipmeijer, *The effect of CEO tweets on the stock market activity,* April 2020.

[14] S. SHIHAB ELBAGIR and Y. JING, "Twitter Sentiment Analysis Based on Ordinal Regression," vol. 7, p. 9, 2019.

[15] G. M. Alec, B. Richa and H. Lei, "Twitter Sentiment Classification using Distant Supervision," vol. 150, no. 12, p. 6, 2009.

[16] T. Duyu, W. Furu, Y. Nan, Z. Ming, L. Ting and Q. Bing, "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification," in *52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, USA, 2014.

[17] X. Li, N. Chen, H. Liu and Y. Zou, "Research on Sentiment Analysis Based on Representation Learning," vol. 55, no. 1, pp. 105-112, 2019.

[18] D. Liben-nowell and J. Kleinberg, "The Link Prediction Problem for Social Networks.," in *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, New Orleans LA USA, 2003.

[19] L. Xinyu and T. Zihan, "Sentiment Analysis on Inflation after COVID-19," 2022.

[20] A. Alsayat, "Improving Sentiment Analysis for Social Media Applications Using an Ensemble Deep Learning Language Model," Arabian Journal for Science and Engineering, 2022.

[21] S. Anubhav, S. Chakraborty, S. Ghosh and S. K. Naskar, "Evaluating Impact of Social Media Posts by Executives on Stock Prices," vol. 2, no. 1, p. 9, 2022.

[22] Z. Lei, R. Ghosh, M. Mohamed and H. Meichun, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis," HP Laboratories , 2011.

[23] N. R, S. S, A. A.M, A. A and N. K. M, "An Ontology based Sentiment Analysis for mobile products using tweets," in *Fifth International Conference on Advanced Computing*, 2013.

[24] S. P and M. I. S. L, "Sentiment analysis of online product reviews using DLMNN and future prediction of online product using IANFIS," p. 20, 2020.

ACCOMPANYING REPORT

The COMP516 Module sets out to provide a deep and systematic understanding of the nature and conduct of Computer Science research.

Our Objective - To prepare a Literature Review on any topic of your choice as a group.

In the very beginning of the semester, we were asked to form groups. The current group has 7 members, they are - 1. Bharathidasan Ilango 2. Giridhar Dhanapal 3. Indrajith Shanmugasundaram 4. Jafreen Kazi 5. Sanjana Tarekar 6. Shreya Krishnarth 7. Sivaramakrishnan Sridharan

Once the group was formed, we were asked to submit the topic. This part was the easiest for us, as in the welcome week when we all met, my team members, the only question after asking ‘Where are you from?’ was ‘Are you on social media?’, ‘Can you please add on LinkedIn?’, ‘Do you use Twitter?’, etc. This is when we realized that if we are so much connected to this virtual world why not try to understand a bit more about it.

The year 2022, is the year when all of us got to see each other in person. This gap of 2 years was tough for all of us. We were so happy to experience the warmth from everyone at the university. Not only us but all on the campus were sharing their stories of how they spent these two years of pandemic. I still remember Bharathi saying that ‘It was not easy, but I am blessed that I got a chance to take care of my family’. This emotion is mixed, it is happy and heart-touching but also, we felt sad for all the tough time we all have been through.

During the time of covid, Sanjana spent time on YouTube, and this was the first time when she learnt about people use sentiment analysis tools in court to address legal matters. One of the most viewed legal trials on YouTube was defamation lawsuit filed against Amber Heard by Johnny Depp. These are the executive people, the celebrities, the influencers who has very deep influence on the sentiment of the general people. Opinion mining using twitter dataset was used for this trail which was considered by the judge as one of the rulings.

Approach to solve a problem has become digital and is now spread worldwide. Siva had a similar interest in football. He is a Ronaldo fan, well, who is not? When Rolando sat on a conference and replaced a coke bottle with water and said ‘Aqua’ which means water in his native language. The Coca-Cola industry saw a brutal drop on the brand value. A company’s stock plummeted with a total loss of 4Billion dollars just by a gesture. Cristiano Ronaldo was clear. "I'm tough on my son. Sometimes he drinks Coke and Fanta, and he eats chips, and he knows I do not like it," he had said.

Countless incidents related to Elon Musk, the most influential CEO, have changed the movement of the stock market. He twitted, ‘You can now buy a Tesla with Bitcoin.’ Indrajith was so interested in cryptocurrency because he invested in it and always wanted to be up to date with this news, articles which affected the movement of the currency.

He also displayed the symbol of dogecoin using drones while the opening of Tesla Gigafactory at Texas. Elon Musk’s announcement on buying twitter created a sudden reflection on stock prices of Twitter and even Tesla. Shortly afterward, Musk, the world's richest person, tweeted that he would "keep supporting Dogecoin" and many activists claimed that he got profited out of that tweet by buying a huge volume of the cryptocurrency before the tweet. Coincidentally, he became the world’s richest person in November 2021 between those tweets.

Most recent incident which shook the entire world, when an innocent young women suffered a severe head injury and was declared dead just for not wearing her Hijab properly. The word ‘Properly’ was emphasized in most of the social media texts. Social media has the potential to unite people globally. People saw what happened and started to raise their voice against injustice. They ran a hashtag ‘#Mahsa_Amini’ and chanted "Woman, Life, Freedom.", threw their hijab in fire and cut their hair as a part of the protest. Power of the electronic world was very clear and lucid that the government had to take serious action against the Mortality Police.

Conversations encounter among us were so intriguing that we considered the topic - ‘**SENTIMENT ANALYSIS IN SOCIAL MEDIA AND FINANCIAL DATA**’.

Brainstorming was next on our agenda. We all sat in the library study room and tried to understand the actual meaning of Sentiment Analysis. We came across several terms like opinion mining, data mining, polarity of the words and many more. These all are directly associated with the various learning algorithms and concepts such as, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Support Vector Machine (SVM), Hybrid model like Multi-Column Convolutional Neural Network (MCNN), Linear Regression (LR), Gradient Boosting Trees (GBT), Bidirectional Encoder Representations from Transformers (BERT), Long Short-Term Memory (LSTM), Ensemble Deep Learning like Bagging, Boosting, Stacking, etc. These all concepts are new for us, but we were aware about them. We started by asking ourselves simple questions such as: -

1. What is a sentiment?
2. How do I analyse the sentiments?
3. What should be the centre of our study?
4. Can we derive a hypothesis?

There was a deluge of heavy and perplexing questions. But, at this point we were determined to study the methods involving in analysing sentiments. So, we added the two most important questions in our Literature Review, they are:

1. How to conduct opinion mining using big data i.e., twitter?
2. Does the sentiment of tweets by executives have a deeper influence than general tweets?

To answer these questions, the amount of study and time spent to understand the methods and algorithm were gigantic. We were compelled to divide our one group into two smaller once.

Shreya, Siva, Jafreen, and Giridhar worked on the first question – ‘How to conduct opinion mining using big data i.e., twitter?’. They considered one main paper for their objective and read more than 23 papers to find a supporting argument. Social media is colossal, but the Found Data also known as Big Data is never ending. They considered numerous algorithms on Machine Learning, traditional approach, various word embedding techniques, several lexicons-based models, customized models and were trying to make the ends meet.

Bharathi, Sanjana and Indrajith worked on second question – ‘Does the sentiment of tweets by executives have a deeper influence than general tweets?’. They used the similar approach as the first group and came up with multiple hypotheses. They read multiple journals, articles even thesis to understand how one influencer can change the mind of millions of people in the world.

Here comes the phase one of the biggest challenge, how to find a common ground on both the questions. The answer is quite simple – Its ‘Big Data’. Data generated on twitter by users is around five hundred million per day. We knew that both the questions can be answered using twitter data, but we also considered data from Reddit, Yelp, Amazon, YouTube, etc. This helped us gain a better insight of the movement of data. Novel-coronavirus is the most recent pandemic event that occurred in the world and the changes it caused was unimaginable. Ongoing pandemic is still affecting people, as we are writing this review. So, most of the papers we considered were published recently. SARS-CoV-2 was also considered to understand the previous data. It contributed to tremendous increase in data volume which allowed improvement in the algorithm to learn for the patterns in the data. For historic financial data we used Yahoo Finance as our primary source. The part where we split groups proved efficient to overcome these challenges.

Phase two was understanding the various concepts used for sentiment analysis. We listed out all the concepts like Decision Tree, Naive Bayes, CNN, RNN, SVM Hybrid model like MCNN, LR, GBT, BERT, LSTM, Ensemble Deep Learning like Bagging, Boosting, Stacking, etc. Now was the time to understand all the related algorithms in determining the polarity or sentiment. We did a thorough study of these topics, classifiers and algorithms and came with an idea of finding the accuracy of all those models. After numerous attempts, we finally prepared an accuracy table from more than twelve papers. This gave us a proper idea on finding the best model to determine an output. In our case, we derived that Deep ensemble learning model gives us the expected results. Ensemble model is a unique model. It combines state-of-the-art method, uses deep learning framework.

It was time to present our presentation, team prepared the slides and wrote transcripts. Everyone represented the group; 20 minutes presentation was followed by a Question & Answer session. We compiled our study to provide a satisfactory output (i.e., analysing the polarity) during our presentation. We found that there were quite a few areas of improvement. Challenges we faced were finding and accessing the related papers that would help us in making a better literature review. In few of the popular papers the performance metrics were missing, which was a challenging task in determining the efficiency of the technique they used in the respective papers. We initially prepared a draft document that was explaining many additional algorithms with twenty pages, but to fit in our 10- page review, we then reduced the optional explanatory part of it. Avoiding plagiarism and ensuring that we gave proper citation for every source considered in the final review was time consuming. It took us more than a week to make our report precise, and with proper formatting.

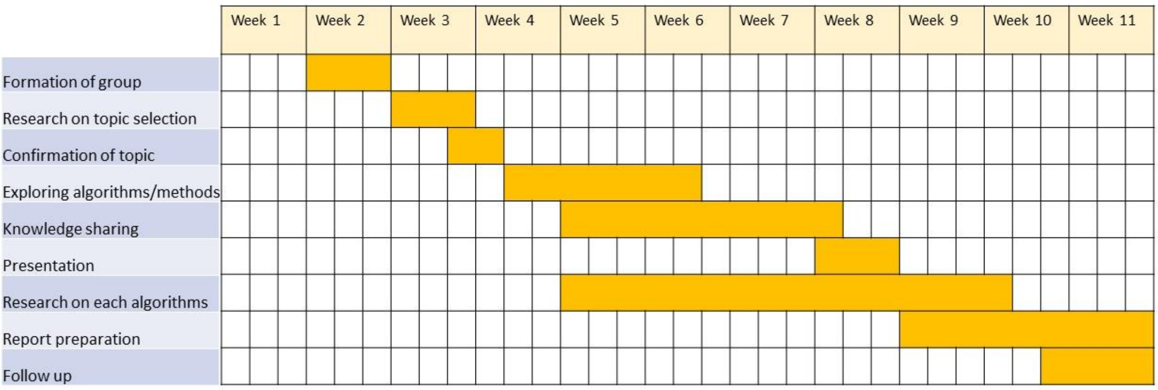

Table 3: Gantt chart of team's performance metrics

As a team, we all worked together in filling the gap from better to best which is mentioned in the above Gantt chart. The entire 3 weeks were spent on collaborating on the respective domains, the team complied the document and prepared a final draft.