Sentiment Analysis on 10-K Financial Reports using Machine Learning Approaches

Conference Paper · November 2021

DOI: 10.1109/ICSET53708.2021.9612552

2 authors, including:

Tan Chye Cheah
University of Nottingham, Malaysia Campus
4 PUBLICATIONS **5** CITATIONS

Sentiment Analysis on 10-K Financial Reports using Machine Learning Approaches

Gim Hoy, Soong
School of Computing
Asia Pacific University of Technology and Innovation
Bukit Jalil, Malaysia
ghsoong@hotmail.com

Dr. Chye Cheah Tan
Faculty of Computing, Engineering, and Technology
Asia Pacific University of Technology and Innovation
Bukit Jalil, Malaysia
chyecheah.t@staffemail.apu.edu.my

***Abstract*— 10K Financial reports are submitted by public listed companies to the Security Exchange Commission (SEC) yearly or quarterly. It allows investors to understand strategic planning and directions of the business organization. Although true facts are required to be presented in the reports, it does not prevent companies from using confusing explanations to beautify the organizations current state. Hence, an automated approach to filter out sentiments from the reports is crucial to assist investors in evaluating financial reports. This research paper explores machine learning approaches to conduct sentiment analysis on 10K financial reports. Two different datasets were intended to be used for training the model but only the financial phrase bank dataset was used to produce the final machine learning models. Four machine learning models including fastText, Naïve Bayes Support Vector Machine (NBSVM), Bidirectional Gated Recurrent Units (BiGRU), and Bidirectional Encoder Representations from Transformers (BERT) are trained based on the financial phrase bank dataset. It is discovered that the BERT model performed with the best accuracy while testing the models while the fastText model provided the fastest loading and training time. Conclusion of this research paper shows that different machine learning models in sentiment analysis possess respective advantages and disadvantages and further research can be done with the combination of textual and numerical data in financial reports.**

Keywords—Machine Learning, Sentiment Analysis, Financial Text Analysis, 10k report.

I. INTRODUCTION

Financial literacy is without a doubt an important field of knowledge to have in this globalized era. Financial reports have always been good to go place for investors who are looking to invest and make decisions on whether the business is profitable. However, not everyone holds the knowledge on the effects of the numbers in a financial report and the stance of the company in the future. Financial reports contain related company information that is disclosed to the public for a better understanding of the company's monetary position in the market. They are essential for investors to evaluate companies for potential investments. 10-K financial reports, also known as Form 10-K, are required to be submitted by most United States public companies to the United States Securities and Exchange Commission (SEC) every year. Compared to the annual report distributed to shareholders, the 10-K reports usually contain more detailed and accurate information due to the laws and commandments by the SEC which forbids misleading and false information. [1]

Text analysis of financial reports has space for improvement due to the lack of models with words specific to the financial field. Financial reports like the 10-K report can also contain words that are deemed useless in a general text analysis model and cause the analysis to be ineffective. A lack of context of the financial report will often lead to a confused understanding of the investor on the current situation of the company. Financial textual data has shown its purpose in predicting foreign exchange prices [2]. Its usage to detect bankruptcy and fraudulent companies [3] can also prevent investors from putting their money into the wrong companies. These efforts have proven that there is much to learn and absorb from company financial reports that can demonstrate a company's existing situation, future and past. Even country currencies have connection with financial textual data. Great Britain pulling out from the European Union [4] is a strong example of events affecting foreign exchange and which the financial new headlines of it contain much more information that investors can investigate. Much research on financial textual data relies on the bag-of-words approach which only counted the number of positive or negative words in the text.

Financial reports can be deceptive as companies try to engulf their organisations with positivity for public image. One financial report can even cause multiple financial experts to be unsure of the sentiments it is holding. The lengthy financial reports also hold knowledge within, but additional words are added to confuse the perceptions towards the company's future [5]. It would be too time-consuming for an investor to fully understand and read through each single sentence in a financial report. With the improvement in computer hardware and natural language processing models, it is recommended that investors utilise these tools to assist in filtering company financial reports. The application of sentiment analysis has been used in various business and social areas and it should not be forgotten in the thriving financial market.

This research intends to utilize the power of machine learning for sentiment analysis on financial textual data. Sentiment analysis can be used to identify if the company has done well or not in the past financial year. The main objective of this research is to review the performance of four different machine learning models which are fastText [6], Naïve Bayes Support Vector Machine (NBSVM) [7], Bidirectional Gated Recurrent Units (BiGRU) [8], and Bidirectional Encoder Representations from Transformers (BERT) [9] in deriving sentiments from 10K financial reports. Investors can now first understand the document's sentiment overview before stepping in to read the financial report thoroughly. A sentiment analysis model can also help investors identify key points and outlooks in a financial report more efficiently before making an investment choice. In the remaining of the paper, section 2 will be a domain

research where the researcher reviews journal articles and conference proceedings related to sentiment analysis on financial text. Previous attempts and research done to build models that can predict sentiments from financial textual data are analysed in the section. Section 3 will talk about the methodology of the research project and section 4 will discuss about the results. The research paper will then end with conclusions in section 5.

II. Sentiment Analysis on Financial Text

A study by Shuhidan [10] showed the use of sentiment analysis on Malaysian financial news headlines. A lexicon-based algorithm and Naïve Bayes algorithm was used to carry out sentiment analysis on financial texts. The study used financial news obtained from the business section of the New Straits Times, a local newspaper which talks about the Malaysian financial economy, for a duration of 12 months. Pre-processing steps to apply the algorithms such as data extraction, stemming and stop words removal were taken in account to prepare the financial textual data. Shuhidan showed the use case of snowball stemmer and R programming language in the process of performing sentiment analysis on financial text.

Atzeni [11] performed a fine-grained approach towards sentiment analysis and obtained true sentiment score values. Several classifiers were used in the approach and lead to an accuracy level of 72%, proving the success their approach in sentiment analysis on financial text. Their dataset consisted of microblog messages that focused on stock market events and financial news headlines scraped from different sources. The research showcased the usage of Random Forest, Linear Regression, Lasso Regression, Ridge Regression and Support Vector Machine to classify the headlines and messages into sentiment scores. The scores obtained showed the usefulness of financial textual data to predict the direction of stock prices.

The approach taken by John [12] to perform sentiment analysis focused more on preparing the financial textual data. The financial text used was converted to numeric vectors through N-gram, Term frequency-inverse document frequency, and Paragraph Vector. Several regression models were then trained using the vectorized data. The paper also exhibits the versatility of Python and its library in building machine learning models. Extra datasets such as downloading whole articles, amazon product reviews and financial phrase bank was used to augment the training dataset, but it was proven that no improvements were achieved. John and Vechtomova conclude the paper that there is a lack of reliable financial textual datasets available for training sentiment analysis models.

According to a deep learning model developed by Sohangir [13], machine learning approaches to sentiment analysis help save lots of time by increasing the layers of processing within the model. Traditional lexicon-based approaches require more manual work and manual tokenizing of the financial words. A machine learning approach such as the convolutional neural networks used enables a window of using n-grams to understand the sentiment of the document. Comparing between different deep learning methods, it was concluded that convolutional network is the best to work with for sentiment analysis of financial textual data.

Another work by Singh [14] shows the implementation of WEKA software for sentiment classification and explores four machine learning classifiers carried on sentiment analysis. It was advised that the researcher define four levels of text classification which are document level, sentence level, word level and character level. A methodology was also proposed for the optimization of data mining methodologies to fit into machine learning approaches. The results of the comparison between Naïve Bayes, J48, OneR and BFTree shows that the OneR exhibits better precision, the Naïve Bayes exhibits faster learning rate, the J48 exhibits better true and false positive rates.

An interesting deep learning model presented by Vargas [15] used a recurrent convolutional neural network to predict stock price movement. The model uses technical indicators and receives news title as input. A relationship between stock price and news title was proven and both news and technical indicators show prominent effects on the stock price. It is however, also stated that the news title only affects the stock price temporarily as input of news title a day before forecasting has a better result than models using news titles from past weeks. Another imminent takeaway from this paper is the effectiveness of using sentence level to detect sentiments rather than word level.

This research paper is aimed to make reading 10K financial reports easier for non-financial experts. The comparison of different machine learning models and their abilities in deriving sentiments of financial reports helps discover each models' advantages and disadvantages. Most of the previous papers mentioned dealt with shorter financial textual data such as headlines and twitter messages. This research paper will focus on interpreting the ability of machine learning models in deriving sentiments from lengthy financial reports.

III. Methodology

Two attempts were made by the researcher to train the machine learning model. The researcher intended to self-annotate 10-K financial reports downloaded from the SEC based on positive and negative words present inside the report. The process of that first attempt is shown in Fig. 1 below.

100 random 10-K financial reports of years 2015 to 2020 were downloaded from the SEC. The metadata such as accession number, filing data and last changing data were filtered out along with the main body of the financial report. The main body was inclusive of HTML tags, so the researcher had to clean the tags before creating a Dataframe to compile both metadata and textual data. The Loughran McDonald dictionary [16] was used to identify positive words and negative words from the body of the financial reports. The percentage of both positive and negative words were compared, and the researcher planned to use these percentages to determine that the report is either positive or negative. After discussion, the researcher regarded this method inaccurate since the number of positive words in the financial report could not determine the sentiment correctly. It was decided that this data set to be more suitable for topic of interest modelling instead of sentiment analysis. The result of this attempt includes downloaded financial reports from 100 companies and an Excel file that contains the percentage of words matched with the words in the Loughran McDonald master dictionary. No machine learning models able to

derive sentiments from 10K financial reports were created in this attempt.

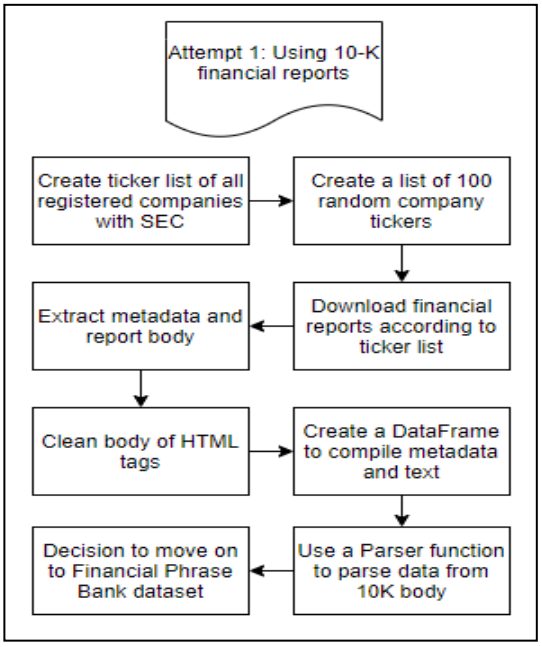

Fig. 1: Flowchart for 10-K financial report attempt.

The researcher then proceeded on with using the financial phrase bank dataset [17] which was professionally annotated by experts in the financial industry. The process of this attempt is shown in Fig. 2 below.

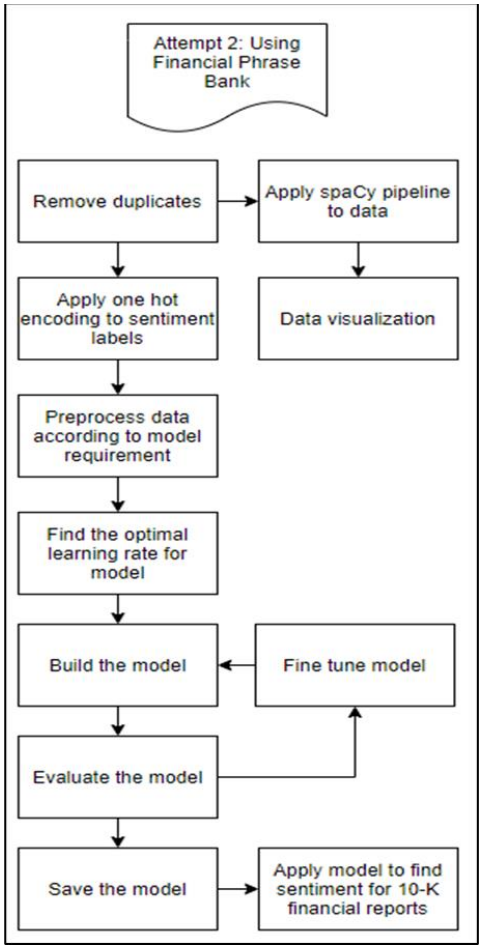

Fig. 2: Flowchart for Financial Phrase Bank Attempt.

Duplicates in the dataset were first eliminated and the data was prepped for data visualization to learn more about the dataset. The spaCy textual data pre-processing pipeline was used to clean the data of numbers. The number of observations in the financial phrase bank used for data visualization and model building in later steps was 4845 rows. The researcher decided to use all rows of observations since natural language processing relies strongly on the amount of data used to train the model.

During the data visualization phase, it was found out that most of the sentences were labeled as neutral with 2872 observations found. There were 1363 sentences with positive sentiments and 604 sentences with negative sentiments. The researcher also discovered that most of the financial new headlines were obtained from Europe financial news headlines, causing the most frequent word appearing to be 'eur' which stands for Europe.

One hot encoding was then applied to the sentiment label column where the labels were differentiated into their respective columns with '0' representing absent and '1' representing present. The whole dataset was also split into two where 30% of the dataset was used as testing dataset and 70% of the dataset was used as training dataset. Four machine learning models were decided to be used which are fastText, Naïve Bayes Support Vector Machine (NBSVM), Bidirectional Gated Recurrent Units (BiGRU), and Bidirectional Encoder Representations from Transformers (BERT). The optimal learning rate for each model were first identified with the automated learning rate finder available in the ktrain library. The researcher then started training the model using the training data. Four models were built, and the optimal learning rate was used to ensure that the models were at their best fit.

To evaluate the model, the accuracy is taken into account. Accuracy of the model is calculated by the number of correct sentiments predicted by the model for the financial news headline. It is used in the results and discussions to identify the most accurate model. The model is built by feeding labelled data where sentiments of financial news headline are provided along with the text itself. For example, a 90% accuracy of a model means the model is able to derive the correct sentiment 90% of the whole test dataset. The value of accuracy will only lie between 0 to 1 which, multiplied by 100, will get the accuracy percentage of the model.

After receiving the results from evaluating the model, the researcher will then fine-tune the hyperparameters by trial and error. The hyperparameters changed includes the learning rate, the batch size, and number of epochs the model will go through. Different values of hyperparameters are tested to find the best fit. Early stopping is used when there is no significant increase in accuracy after two epochs, where the model will stop training even if it has not achieved the set number of epochs. Weights from the best epoch is then loaded into the model and the model with the best accuracy is saved.

This paper intends to derive sentiments from 10-K financial reports. Hence, the models built will be used in predicting sentiments from 10-K financial reports of Apple Inc. from the year 2015-2020. The results are presented in the next section along with discussions derived.

IV. RESULTS AND DISCUSSION

A. Attempt 1 10-K financial report result

In the attempt, a total of 572 financial reports were pre-processed and the number of words, positive words percentage, negative words percentage, uncertainty words percentage, litigious words percentage, weak-model, moderate-model, strong-modal, and constraining words percentage from each financial report were identified. It was discovered that the percentage of positive words is obtained by dividing the number of matched positive words with the Loughran McDonald dictionary over the total words in the report and it is likewise for the percentage of other categories. The number of positive words can be deduced by multiplying the percentage with 100 and number of total words in the report. The researcher suggests that using the matched positive words or negative words is not convincing enough to indicate the sentiment of the whole financial report as there are words not within the dictionary filtered out. Hence, it is concluded that only looking at the percentage of positive words or negative words percentage to determine the sentiment of the financial report is not suitable for training machine learning models as there are a lot of other noise in the data. It is recommended that this attempt is used on topic modelling instead of sentiment analysis. A portion of the results produced in attempt one is shown in Fig.3 below.

number of words,	% positive,	% negative,	% uncertainty,	% litigious,	% modal-weak,	% modal moderate,	% modal strong,	% constraining,
87802	0.402040956	1.161704745	1.050089975	0.764219494	0.789275871	0.202728867	0.437347669	0.496571832
186554	0.30286137	0.850692025	0.544614428	0.544614428	0.249793625	0.137225683	0.154378893	0.263730609
1993928	0.61451567	0.24674913	0.916733202	0.180798905	0.029389226	0.019208316	0.023471259	0.067755706
1930759	0.702728823	0.246483378	0.957809856	0.171538758	0.025275034	0.017868621	0.019059862	0.06375731
1963967	0.692832415	0.248884019	0.943243955	0.184524485	0.02943023	0.017057313	0.023676569	0.065937971
1836129	0.727998959	0.246870079	0.944105779	0.200149336	0.029954322	0.018789529	0.028592762	0.063721813
1751356	0.67473432	0.230621301	0.951205809	0.196704725	0.032260717	0.02141198	0.02923449	0.077768312
1872227	0.653927115	0.28650372	0.910199458	0.282497795	0.039685359	0.025637917	0.036694268	0.087542803

Fig. 3: Financial report word contents

B. Attempt 2 Financial Phrase Bank result

All sentiment analysis models which include fastText, Naïve Bayes Support Vector Machine (NBSVM), Bidirectional Gated Recurrent Units (BiGRU), and Bidirectional Encoder Representations from Transformers (BERT) were loaded inclusive of their pre-processing pipelines. The time required to load each model was recorded as shown in Fig. 4 below.

```
NBSVM model loaded complete in  0:00:00.744862
fastText model loaded complete in  0:00:00.152135
BiGRU model loaded complete in  0:00:00.585517
BERT model loaded complete in  0:00:04.910042
```

Fig. 4: Time taken to load each model.

Each model is then used to perform sentiment analysis on 10-K financial reports of Apple from year 2015 to 2020. The time required for each model to predict sentiments and their results are then recorded as well as shown from Fig. 5 to Fig. 8.

```
NBSVM Model Prediction start
2015  Apple 10-K Financial Report negative
2016  Apple 10-K Financial Report negative
2017  Apple 10-K Financial Report negative
2018  Apple 10-K Financial Report negative
2019  Apple 10-K Financial Report negative
2020  Apple 10-K Financial Report negative
Total time used:  0:00:08.378868
```

Fig. 5: Time and results for NBSVM model predictions.

```
fastText Model Prediction start
2015  Apple 10-K Financial Report positive
2016  Apple 10-K Financial Report neutral
2017  Apple 10-K Financial Report positive
2018  Apple 10-K Financial Report positive
2019  Apple 10-K Financial Report neutral
2020  Apple 10-K Financial Report positive
Total time used:  0:00:04.457226
```

Fig. 6: Time and results for fastText model predictions.

```
BiGRU Model Prediction start
2015  Apple 10-K Financial Report neutral
2016  Apple 10-K Financial Report neutral
2017  Apple 10-K Financial Report neutral
2018  Apple 10-K Financial Report neutral
2019  Apple 10-K Financial Report neutral
2020  Apple 10-K Financial Report neutral
Total time used:  0:00:05.421981
```

Figure 7: Time and results for BiGRU model predictions

```
BERT Model Prediction start
2015  Apple 10-K Financial Report neutral
2016  Apple 10-K Financial Report neutral
2017  Apple 10-K Financial Report neutral
2018  Apple 10-K Financial Report neutral
2019  Apple 10-K Financial Report neutral
2020  Apple 10-K Financial Report neutral
Total time used:  0:04:11.980732
```

Fig. 8: Time and results for BERT model predictions.

The NBSVM model assumed all the apple financial reports to have a negative sentiment whereas the BiGRU and BERT model assumed that all financial reports carry a neutral sentiment. The fastText model predicts that reports from 2016 and 2019 to have a neutral sentiment and other years brought a positive sentiment. From this result, the researcher remembered that most of the data in the dataset was annotated as neutral. Since the BERT model and BiGRU model relies heavily on a unbiased dataset, a biased sentiment analysis may be performed in result to the reports being all detected as neutral. The NBSVM model predicting that all financial reports imposing a negative sentiment may be due to various occurrences of negative word vectors in the financial reports. Lastly, the fastText model is reported to use averaging on word features to form good sentence representations and hence the positivity in the sentiment analysis. The time used for the BERT model is caused by the complexity of the model which required not only longer time to train and load but to predict sentiments as well.

TABLE I. COMPARISON OF RESULTS

Model	Training Time	Loading Time	Prediction Time	Accuracy
NBSVM	7 seconds	0.74 seconds	8.4 seconds	72.83%
fastText	41 seconds	0.15 seconds	4.5 seconds	78.75%
BiGRU	58 seconds	0.59 seconds	5.4 seconds	78.82%
BERT	1.03 hour	4.91 seconds	4 minutes 12 seconds	90.08%

Table 1 above shows how each model performs in contrast of one another. The BERT model has the best

accuracy amongst all models. However, the downside of the BERT model is the longer training, loading, and prediction time required to perform sentiment analysis. It is advised that the BERT model be used in situations where accuracy is more important than speed. The complexity of the model is one disadvantage and the reason the model requires longer time for training, loading and prediction. The performance of the NBSVM model is the worst amongst all four models with low accuracy. The training time for the model is the only advantage the model possesses and does not do exceptionally well in loading and prediction time. The BiGRU model has similar performance with the fastText model with almost equal accuracy. The advantage of the fastText model is the time required for loading and prediction. The training time of the fastText model is relatively low compared to BiGRU and BERT models. The shorter time needed to load and predict makes the fastText model suitable for on-site sentiment analysis. In summary, the fastText model can be chosen to analyse large number of financial reports in a single go where the BERT model can be chosen to obtain a more accurate sentiment analysis.

V. Conclusion

Data pre-processing steps to ensure that the data is kept within the acceptable form of input allows the machine learning models to be built and make predictions. The financial export annotated financial phrase bank dataset was proven suitable for supervised machine learning approaches to sentiment analysis. All machine learning models were able to predict sentiments with their respective data pre-processing pipelines. The results of using these models to derive sentiments of Apple company from 2015 to 2020 gave an interesting insight in the advantages and disadvantages in each model.

After comparing the results and evaluation, the researcher concludes that the Bidirectional Encoder Representations from Transformers (BERT) model has the best performance in terms of accuracy. The model is able to predict the right sentiments 90% of the time. However, the model is not its best in loading and prediction time due to its complexity. The fastText model, on the other hand, shows great response to having the shortest loading and prediction time. Although the lower 10% in accuracy reduces its ability to predict the right sentiments, the fast prediction time allows more predictions in a single amount of time and makes the model suitable for going through large number of financial reports.

The researcher recommends future research to be done on combining the numerical data available in the financial report with the financial textual data to provide a more thorough view of the whole financial report. More datasets like the financial phrase bank are encouraged to be explored and future research can be done on comparing the results of sentiment analysed by models trained on different datasets. Future research can also be done on aspect-level sentiment analysis on financial reports instead of the document-level sentiment analysis done by the researcher. It is suggested to continue working on a topic modelling path to extract topics of interest in financial reports.

Acknowledgment

I would like to express my sincere gratitude to my supervisor, Dr. Tan Chye Cheah for the continuous support of the research project, for his patience, encouragement, knowledge, and advice. Finally, I would like to thank my parents for bringing me to this world and providing all the support they could give throughout my life.

This work was supported by the APU Faculty Research Grant scheme, project number: APURDG/01/2019.

References

[1] U.S. Securities and Exchange Commision, "How to Read a 10-K," 2011. [Online]. Available: https://www.sec.gov/fast-answers/answersreada10khtm.html. [Accessed 28 November 2020].

[2] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah and D. C. L. Ngo, "Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment," *Expert Systems with Applications*, vol. 42, no. 1, pp. 306-324, 2015.

[3] M. Cecchini, H. Aytug, G. J. Koehler and P. Pathak, "Making words work: Using financial text as a predictor of financial events," *Decision Support Systems*, vol. 50, no. 1, pp. 164-175, 2010.

[4] IG, "The value of the pound since Brexit," 2021. [Online]. Available: https://www.ig.com/en/financial-events/brexit/value-of-the-pound-since-brexit. [Accessed 15 August 2021].

[5] B. Wang and X. Wang, "Deceptive Financial Reporting Detection: A Hierarchical Clustering Approach Based on Linguistic Features," *Procedia Engineering*, vol. 29, pp. 3392-3396, 2012.

[6] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, "Bag of Tricks for Efficient Text Classification," 2016. [Online]. Available: arXiv:1607.01759v3.

[7] S. Wang and C. Manning, "Baselines and Bigrams: Simple, Good Sentiment and Topic Classification," in *50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, 2012.

[8] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch and A. Joulin, "Advances in Pre-Training Distributed Word Representations," 2017. [Online]. Available: arXiv:1712.09405v1.

[9] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019. [Online]. Available: arXiv:1810.04805v2.

[10] S. M. Shuhidan, S. R. Hamidi, S. Kazemian, S. M. Shuhidan and M. A. Ismail, "Sentiment Analysis for Financial News Headlines using Machine Learning Algorithm," Kuching, Sarawak, Malaysia, 2018.

[11] M. Atzeni, A. Dridi and D. R. Recupero, "Fine-Grained Sentiment Analysis on Financial Microblogs and News Headlines," in *Semantic Web Challenges*, vol. 769, D. M., S. M. and B. E., Eds., Heraklion, Crete, Greece, Springer, 2017, pp. 124-128.

[12] V. John and O. Vechtomova, "Sentiment Analysis on

Financial News Headlines using Training Dataset Augmentation," 2017.

[13] S. Sohangir, D. Wang, A. Pomeranets and T. M. Khoshgoftaar, "Big Data: Deep Learning for financial sentiment analysis," *Journal of Big Data,* vol. 5, no. 3, 2018.

[14] J. Singh, G. Singh and R. Singh, "Optimization of sentiment analysis using machine learning classifiers," *Human-centric Computing and Information Sciences,* vol. 7, no. 32, 2017.

[15] M. R. Vargas, B. S. L. P. d. Lima and A. G. Evsukoff, "Deep learning for stock market prediction from financial news articles," in *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, Annecy, 2017.

[16] T. Loughran and B. McDonald, "Textual Analysis in Finance," *Annual Review of Financial Economics,* vol. 12, pp. 357-375, 2020.

[17] P. Malo, A. Sinha, P. Korhonen, J. Wallenius and P. Takala, "Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts," *Journal of the American Society for Information Science and Technology,* vol. 4, no. 65, pp. 782-796, 2014.