Hate Speech in Filipino Election-Related Tweets: A Sentiment Analysis Using Convolutional Neural Networks

Thesis · October 2024

DOI: 10.13140/RG.2.2.20961.52326

Charmaine Ponay
University of Santo Tomas

Hate Speech in Filipino Election-Related Tweets: A Sentiment Analysis Using Convolutional Neural Networks

Sharmaine Crystel D. Argañosa
Department of Computer Science
College of Information and Computing Sciences
University of Santo Tomas, Manila, Philippines
sharmainecrystel.arganosa@gmail.com

John Jeco S. Villanueva
Department of Computer Science
College of Information and Computing Sciences
University of Santo Tomas, Manila, Philippines
johnjeco.villanueva@gmail.com

Raphaelle L. Marasigan
Department of Computer Science
College of Information and Computing Sciences
University of Santo Tomas, Manila, Philippines
raphaellemarasigan@gmail.com

Joan Kirby C. Wenceslao
Department of Computer Science
College of Information and Computing Sciences
University of Santo Tomas, Manila, Philippines
joankirby.wenceslao@gmail.com

Charmaine S. Ponay
Department of Computer Science
College of Information and Computing Sciences
University of Santo Tomas, Manila, Philippines
csponay@ust.edu.ph

***Abstract*—The use of social media applications has served as a platform for communication of information and expression of opinion. Twitter, one of the most commonly used social media applications, opened online interactions that allowed users to express their opinions on different topics, including politics. Tweets publicized during the 2022 Philippine national elections were gathered and classified as either positive, negative, or neutral and were categorized as either hate or non-hate using Convolutional Neural Networks. The labeled dataset was balanced to have similar quantities per class to fit the ratio splits 70:30, 80:20, and 90:10. The proposed model, fastText CNN, had an overall better performance in comparison to the reference study, TF-IDF Feedforward Neural Network. Particularly, among the three splits with two models the study has conducted, the fastText CNN binary classification model with a 90:10 split performed best with an accuracy of 83.79%, precision of 83.97%, recall of 83.79%, F-measure of 83.77%, and an average runtime of 0.0474 seconds.**

Keywords: *Bilingual, Machine Learning, Deep Learning, Natural Language Processing, Sentiment Analysis, Convolutional Neural Network, fastText Word Embeddings, Feedforward Neural Network, TF-IDF, Hate Speech*

I. Introduction

In today's growing generation centered on technology, social media applications such as Facebook, Youtube, and Twitter are commonly used by internet users. Among the different social media applications, Twitter was utilized as a way for registered people to connect and disseminate information. It was also used for keeping up with the news, marketing campaigns, and entertainment, allowing users to express themselves through likes, retweets, and tweets freely. The online platform's accessibility and ease of use allowed users to utilize their freedom of expression and interact with other users, which is part of its popularity as a social media application.

The Philippines was considered the social media capital of the world [1] and was also known as one of the top countries that use Twitter. The increase in social interactions on Twitter was not limited to discourse, as this had also made room for tweets containing offensive words against individuals. The significance of a social media network's speed, accessibility, and permission to anonymity had contributed to problems concerning hate speech [2] which was described as a speech or writing that was bias-motivated, hostile, or malicious that conveyed prejudice aimed at a person or group based on characteristics such as gender, race, religion, ethnicity, color, national origin, disability, or sexual orientation [3].

Computer science has advanced and brought forth applications that can be used for real-world problems, such as hate speech in social media can be in the form of tweets.

The researchers aimed to utilize Convolutional Neural Networks for the sentiment analysis of Filipino election-related tweets into the classification of positive, neutral, and negative and are categorized into hate or non-hate using tweets from the 2022 national elections campaign periods. The use of CNN addressed the hope of maximizing the accuracy results in determining hate speech by utilizing frameworks for text analysis with deep learning features and capabilities.

II. Related Literature

A. Natural Language Processing

Natural Language Processing (NLP) is the study and development of computer systems to have the ability to make interpretations of speech and text just as how human language is spoken and written. The vast and continuous development of human communication led to ambiguity due to colloquialisms, abbreviations, misspellings, and other sorts of irregularities but are nonetheless understood by people. The unpredictability of human language has made it difficult to gain insight into the natural language [4], but with the progression of NLP by combining computational linguistics with statistical, machine learning, and deep learning models [5], understanding the complexity of human language and its sentiments has been made possible. [6] stated that natural language processing is a collection of computational techniques that allow automatic analysis and representations of human languages. Some examples of how NLP is utilized include online information retrieval, aggregation, classification of text into categories, automatic language translation, and question-answering, which are based on algorithms that rely on textual data.

B. Machine Learning

Machine learning became a prominent approach to automated tasks. It provides objectivity based on patterns learned from past experiences and efficiency in which the model improves performance as long as its training data does not overfit. The requirements to build a machine learning model come from creating a training data dataset. This is made to make classifications, predictions, and decisions without explicitly programming the algorithm to perform a specific task. There is a wide range of machine learning algorithms, and as of the date of writing, there are five known types: supervised, semi-supervised, unsupervised, active, and reinforced learning [7]. The type of machine learning used for this study is supervised learning to which text classification algorithms belong. The prerequisite in supervised learning is using labeled or annotated data. However, machine learning algorithms do not read raw text data; for it to read text, the data must undergo text processing which turns the text into vectors for it to be read by the algorithm.

C. TF-IDF

A popular NLP algorithm for the vectorization of text is the Term Frequency-Inverse Document Frequency or TF-IDF which is a technique used to find the meaning of sentences consisting of words and also cancels out the incapabilities of the Bag of Words technique which is helpful for machines to read words in numbers [8]. According to the study of [9], TF-IDF's performance heavily relies on large amounts of training sets for it to have consistent and stable performance, else it would need to be extracted and updated frequently to achieve maximum performance, as compared to fastText which is not necessary. For this study, the data that is used for the machine learning model is pre-processed using multiple layered techniques to attain feasible and usable quality data. Once completed, the algorithm chosen for classification is trained and tested according to its split [10]. Deep learning algorithms are further advancements in machine learning that use neural networks.

D. Convolutional Neural Networks

A related study by [11] conducted the use of Convolutional Neural Networks to classify Hate-Speech found on English Twitter, albeit with the use of a different dataset that one expert and three amateur annotators had annotated. The goal of the CNN classifier was to assign a specific tweet to one of the four designated categories, namely racism, sexism, both, and non-hate-speech. The initial step conducted was to generate feature embeddings. This was done with the processes called word embeddings and character n-grams. As for its implementation, five approaches of CNN using different feature embeddings were tested and these were: random vectors, word2vec, character n-grams, word2vec + character n-grams, and logistic regression with character n-grams developed by the researchers. The study concluded that the best-performing system setup was word2vec alone which accomplished a 78.29% F1-score, the highest overall. F1-score is defined as the harmonic mean of precision and recall. Moreover, the results obtained in the related study can be augmented using CNN without having to test other cases of feature embeddings.

Another related study in the use of CNN authored by [12] published the report "Natural Language Processing for the Identification of Silent Brain Infarcts From Neuroimaging Reports" which aimed to develop NLP systems to determine individuals with incidentally discovered silent brain infarction defined as the presence of 1 or more brain lesions and additionally determining white matter disease which is a common finding in neuroimaging of elderly. Inputs are captured from neuroimaging reports which contain interpretation and findings from neuroimages in unstructured text. The main machine learning algorithm is the CNN implemented using TensorFlow 1.1.02. A rule-based approach is also integrated as an additional NLP approach by using an information extraction task where words curated by domain experts are subjected to classification. The McNemar test is used to evaluate the performance difference between the rule-based and machine-learning models.

E. fastText

A study by [13] used Convolutional Neural Networks for sentiment analysis wherein the words from the dataset are first converted into vectors using word2vec, a conversion proposed by Google to computer vector representation. The vectors reflected the distance of the words and were used to initialize the parameters for the CNN. Their proposed CNN model resulted in an accuracy of 45.4% for their dataset, which performed better than other neural network models such as Recurrent Neural Network (RNN). Another study by [14] also used CNN with fastText Embeddings, wherein the purpose of the word embeddings was to consider the internal structure of words, allowing the representation of rich language. Another study by [15] that utilized fastText word embeddings with their Convolution Neural Network model used Convolutional Neural Networks (for sentiment analysis wherein they stated that the use of fastText with neural networks is a promising approach for text classification. They also stated that the model they proposed was simple and efficient as it only contained three layers but could still give promising results for text classification.

III. Simulations and Results

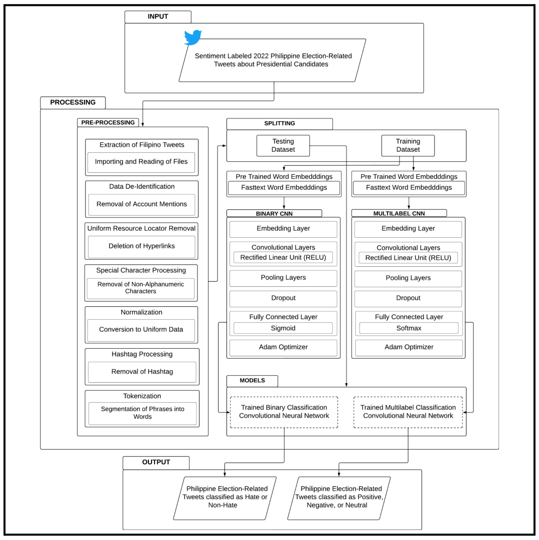

Figure 1. System Architecture

A. *Hypothesis*

This study conducted sentiment analysis on election-related tweets and classified them into hate or non-hate using Convolutional Neural Networks. A study by [2] was used as the basis since their study used Feedforward Neural Networks as the sole model for the deep learning approach, with its model's performance compared with the proposed algorithm Convolutional Neural Networks. In line with this, the researchers formulated the following hypotheses:

Ho: There is no significant difference in the performance between Convolutional Neural Networks and Feedforward Neural Networks in the sentiment analysis of Filipino election-related tweets into hate or non-hate classifications.

Ha: There is a significant improvement in the performance between Convolutional Neural Networks and Feedforward Neural Networks in the sentiment analysis of Filipino election-related tweets into hate or non-hate classifications.

B. *Data Gathering and Preprocessing*

The initial dataset contained 20,000 tweets from different Twitter accounts located in the Philippines whose tweets include the presidential candidates for the 2022 Philippine national elections, dated from the 8th of October 2021 until the 7th of May 2022. Following a set of criteria formulated and based upon, the dataset underwent manual labeling according to their respective sentiment (positive, negative, or neutral), grouped into hate or non-hate, removal of duplicate tweets and languages of tweets outside the scope, leaving a total of 19,012 tweets with 6,162 positive, 10,287 neutral, and 2,563 negative tweets. An instructor from UST Department of Political Science validated the criteria for annotation and the labeled dataset to ensure the integrity of the data used for the experiment. The tweets went through a series of phases for preprocessing, namely: data de-identification where "@" mentions included in the tweet are removed using a regex, URL removal where hyperlinks referring to media from tweets are removed, special character processing where emojis and non-alphanumeric characters were removed using regex, normalization where the tweet was lowercase and removed of stopwords, hashtag processing where hashtags in the tweet were removed, tokenization where tweets were segmented in to words for its use in vectorization for the use of the models. This resulted in two types of datasets: 1) binary classification dataset, which contained 5,120 tweets, and 2) multi label classification dataset, which contained 7,680 tweets. The study followed a 70:30 split but also conducted on 80:20 and 90:10 for additional test cases..

C. *Training, Validation, and Testing*

The tweets datasets were proportioned into three different ratios namely: 70% for training with 30% for testing, 80% for training with 20% for testing, and 90% for training with 10% for testing. After training the model, the testing test set was used for the trained CNN models to execute the sentiment analysis for Filipino election-related tweets in order to determine if the sentiments fall under the classifications of positive, negative, and neutral for the multi label classification models and into hate or non-hate for the binary classification models. The classification results were used for the performance evaluation of the model to compute for accuracy, precision, recall, and F-measure. Testing the model's accuracy was done by comparing the result with the results of the TF-IDF Feedforward Neural Network models from [2] which were also implemented by the researchers following the architecture of the reference study..

The architecture of the proposed models for the binary and multi label classification fastText CNN were based from the study of [6] wherein the kernel and filter sizes were experimented by the researchers until it reached an acceptable and consistent performance. The first convolutional filter had a filter size of 256 and kernel size of 2, and the second convolutional filter had a filter size of 64 and kernel size of 4. The architecture also included an adam optimizer that utilized the default values from the keras package α = 1 × 10-3, β1 = 0.9, β2 = 0.999, ε = 1 × 10-8, and lastly the output layers had a sigmoid function for the binary classification model, and a softmax layer with 3 output nodes for the multi label classification model. The models were then trained within 5 epochs and with the validation split of 0.1 for training. After training, the models were then saved for the conduction of testing wherein the test data were classified by the trained models for the analysis of results.

D. *Results*

Table 1. Performance Measures for fastText CNN and TF-IDF FFNN models for Binary Classification

Split	Accuracy		Precision		Recall		F-measure	
	fastText CNN	TF-IDF FFNN	fastText CNN	TF-IDF FFNN	fastText CNN	TF-IDF FFNN	fastText CNN	TF-IDF FFNN
70:30	**82.62%**	80.08%	**82.72%**	80.17%	**82.65%**	80.11%	**82.61%**	80.07%
80:20	**83.30%**	80.37%	**83.32%**	80.40%	**83.27%**	80.40%	**83.28%**	80.37%
90:10	**83.79%**	83.40%	**83.97%**	83.42%	**83.79%**	83.40%	**83.77%**	83.40%

Table 2. Performance Measures for fastText CNN and TF-IDF FFNN models for Multi Label Classification

Split	Accuracy		Precision		Recall		F-measure	
	fastText CNN	TF-IDF FFNN	fastText CNN	TF-IDF FFNN	fastText CNN	TF-IDF FFNN	fastText CNN	TF-IDF FFNN
70:30	**62.50%**	62.33%	**64.19%**	62.78%	**62.44%**	62.28%	**62.77%**	62.47%
80:20	**64.13%**	63.80%	**64.68%**	63.77%	**64.20%**	63.93%	**64.38%**	63.83%
90:10	**65.23%**	63.02%	**65.59%**	64.52%	**65.46%**	64.02%	**65.44%**	63.06%

The performance of the binary classification fastText CNN models in terms of accuracy, precision, recall, and F-measure show that as the split increases, the performance of the model also increases in terms of the different measures. The best model, according to the results was the binary classification fastText CNN with the 90:10 split that garnered an accuracy of 83.70%, precision of 83.97%, recall of 83.79%, and F-measure of 83.77% which shows that the model with the most amount of training data also showed the best performance, however, the different splits also performed well considering that they were able to result with a greater than 80% accuracy which according to [16] is a good baseline accuracy for models that deal with sentiment analysis. The multi label classification fastText CNN with the 90:10 split that garnered an accuracy of 65.23%, precision of 65.59%, recall of 65.46%, and F-measure of 65.44%, which shows that the model with the most amount of training data also showed the best performance. It is seen that even if the same architecture of the CNN was used as the binary models there is a drop in accuracy, which also follows the same pattern from the reference study by [2] wherein their binary classification models had a higher accuracy than their multi label classification model, which was also similar to the study of [14] wherein their binary classification models also performed better than multi label for sentiment analysis.

Table 3. McNemar's Test with Bonferroni-Holm Method for fastText CNN and TF-IDF FFNN models for Binary Classification

Split	Hypothesis Testing ($\alpha = 0.05$)		
	p-value	HB correction	result
70:30	0.0194	0.025	Reject Ho
80:20	0.0043	0.016	Reject Ho
90:10	0.9062	0.05	Support Ho

The results of the performance measures of the fastText CNN model compared to TF-IDF FFNN using McNemar's test with Bonferroni-Holm correction as statistical test produced the following results. Due to the split of 90:10 for training and testing data, there was no significant difference between the two models as the p-value was greater than the alpha of 0.05 and Bonferroni-Holm correction of 0.05. The performance measures also only had less than 1% difference. However, the results show that lower training splits, such as 70:30 and 80:20 ratios, presented results that rejected the null hypothesis. Therefore, the results show that there was a significant improvement in those splits for the fastText CNN model in comparison to the TF-IDF FFNN model which was also supported by the difference of their performance measures which were at 2-3%.

Table 4. Comparison of Runtime for fastText CNN and TF-IDF FFNN models for Binary Classification

Split	Average Runtime on Test Data	
	fastText CNN	TF-IDF FFNN
70:30	0.0429 seconds	0.0454 seconds
80:20	0.0444 seconds	0.04754 seconds
90:10	0.0474 seconds	0.0507 seconds

The difference in terms of runtime between the two models for binary classification results show almost no difference in the runtime between the two binary models of fastText CNN and TF-IDF FFNN. This is due to the use of the same library for creating the models wherein the only difference is the architecture. Based on the results for the runtime of the TF-IDF FFNN model and the fastText CNN model, the differences are negligible since each prediction only takes less than a second.

IV. Conclusions and Recommendations

A. *Conclusions*

The binary classification fastText CNN models performed well in all performance measures in classifying Filipino election-related tweets into hate or non-hate and achieved measures of greater than 80%.

The proposed fastText CNN outperformed the TF-IDF FFNN model in all performance measures of accuracy, precision, recall, and F-measure of the 70:30, 80:20, and 90:10 splits. The McNemar's test for hypothesis for the performance of fastText CNN, in contrast to, TF-IDF FFNN, exhibited a significant improvement with the 70:30 and 80:20 splits with an improvement of 2-3% in terms of the performance measures. Meanwhile, there is no significant difference for the 90:10 split which had less than 1% difference in the performance measures.

The performances of TF-IDF FFNN and fastText CNN in the average runtime in classifying tweets ranged from 0.04 to 0.05 seconds per tweet, showing that in terms of the models' runtime or efficiency, the differences are negligible since the tweets were classified in less than a second.

The implementation of the proposed system that aims to improve the base study demonstrated that a deep learning model using CNN and fastText word embeddings is comparable to the base model and can be used to classify Philippine election-related tweets into hate or non-hate, and can assist platforms such as Twitter to classify the sentiment of tweets which may be written using the Filipino language.

B. Recommendations

Based on the findings and conclusions of the study, the researchers have recommended the following:

Construction of the convolutional neural network requires a mix of multiple layers that transform the input into a verifiable output using different convolutional, pooling, and connected layers. Often the hyperparameters corresponding to these layers are determined heuristically by previous researchers or tested using loosely defined rules through trial and error. The researchers recommend that further improvements on the hyperparameters and layers used are possible as it was observed from the tests of the researchers that hyperparameters such as kernel sizes and filter sizes of the convolutional and fully connected layers are subject to change as variations in values can impact the model's effectiveness and the output label's accuracy compared to the true label of the data.

The researchers recommend that a greater collection of tweets annotated by more scores of people should be sampled and gathered as it can increase the model's learning capacity since it can discern more token relationships and classification patterns present in each label. More tweets accumulated would mean that the model could benefit from the wider context since it fits the labeling of various tweets in general.

The researchers also suggest that future researchers experiment on not completely removing the hashtags since some contain words or statements that can contribute to the classification of a tweet, specifically those that cause a conflict of labeling sentiments between the tweet and the hashtag. Hashtags may have a representational value when it comes to classifying tweets based on sentiment analysis, as the applied tags are part of the user input in the Twitter platform.

Acknowledgement

The researchers would like to recognize and express their heartfelt gratitude to the following people who have made the success of the study possible.

First and foremost, to the Almighty Lord God, who graced the researchers with strength, intellect, and capability to face drawbacks and proceed successfully with their study.

To Asst. Prof. Charmaine S. Ponay, their technical adviser, for her acceptance to lend guidance to the researchers from the beginning of the study up to the completion of it. The researchers are grateful to her shared insights that made the study successful.

To Asst. Prof. Jannette E. Sideño, their thesis coordinator, for her utmost understanding and patience during the entirety of the study. The study would not be complete without her guidance and motivation.

To Asst. Prof. Donata D. Acula, PhD, Asst. Prof. Cherry Rose R. Estabillo, and Asst. Prof. Darlene D. Alberto, the thesis defense panelists, for their guidance and profound questions given during the defense. Their comments have allowed the researchers to open more perspectives in their study and be of standard to the future researchers.

To Mr. Anthony Andrew G. Divinagracia, MA, who validated the dataset used for the study, hence proving the legitimacy of the study. His ability to corroborate the dataset allowed the researchers to proceed with their study.

Last but not least, the researchers' express their appreciation to their family and friends. They served as the foundation for the researchers' to be kept encouraged, confident, and brave to face both adversities and success in both the performance of the researchers and the completion of the study.

References

[1] Velasco, J. C. (2020). Millennials as digital natives: examining the social media activities of the Philippine Y-generation. Pertanika Journal of Social Sciences and Humanities, 28(3), 1939-1957. Retrieved from http://www.pertanika.upm.edu.my/res ources/files/Pertanika%20PAPERS/JSSH%20Vol.%202 8%20(3)%20Sep.%202020/19%20JSSH-5841-2020.pdf

[2] Cabasag, N., Chan, V. R., Lim, S. C., Gonzales, M. E., & Cheng, C. (2019). Hate speech in Philippine election-related tweets: Automatic detection and classification using natural language processing. Philippine Computing Journal Dedicated Issue on Natural Language Processing, 14(1), 1-14. Retrieved from https://pcj.csp.org.ph/index.php/pcj/issue/down load/29/PCJ%20V14%20N1%20pp1- 14%202019

[3] Cohen-Almagor, R. (2014). Countering hate on the Internet. JRE, 22, 431. Retrieved from papers.ssrn.com/sol3/papers.cfm?abstract_id=2543511.

[4] Barba, P. (2020, September 29). Machine Learning (ML) for Natural Language Processing (NLP). Lexalytics. Retrieved from https://www.lexalytics.com/lexablog/machine-learning-natural-language-processing#background-nlp

[5] IBM Cloud Education. (2020, July 2). What is Natural Language Processing? Www.ibm.com. Retrieved from https://www.ibm.com/cloud/learn/ natural-language-processing

[6] Chowdhary, K. R. (2020). Fundamentals of artificial intelligence. New Delhi: Springer India. Retrieved from https://link.springer.com/content/pdf/10.1007/ 978-81-322-3972-7.pdf

[7] Lytvyn, V., Sharonova, N., Hamon, T., Vysotska, V., Grabar, N., & Kowalska-Styczen, A. (2018). Computational linguistics and intelligent systems. In CEUR workshop proceedings.

[8] Madan, R. (2019, November 27). TF-IDF/Term Frequency Technique: Easiest explanation for Text classification in NLP with Python. Medium. https://medium.com/analytics-vidhya/tf-idf-term-freque ncy-technique-easiest-explanation-for-text-classificatio n-in-nlp-with-code-8ca3912e58c3

[9] van Tussenbroek, T., Viering, T., Makrodimitris, S., Naseri Jahfari, A., Tax, D., & Loog, M. (2020). Who said that? Comparing performance of TF-IDF and fastText to identify authorship of short sentences. Repository.tudelft.nl.https://repository.tudelft.nl/islando ra/object/uuid:93873bbf-2886-4023-b696-e11be2b9902 4

[10] Masch, C. (2021). GitHub - cmasch/cnn-text-classification: Text classification with

Convolutional Neural Networks on Yelp, IMDB & sentence polarity dataset v1.0. GitHub. Retrieved from https://github.com/cmasch/cnn-text-classification

[11] Gambäck, B., & Sikdar, U. K. (2017). Using Convolutional Neural Networks to Classify Hate-Speech. Proceedings of the First Workshop on Abusive Language Online, pages 85–90. Retrieved from http://aclanthology.lst.uni-saarland.de/W17-3013.pdf

[12] Fu, S., Leung, L. Y., Wang, Y., Raulli, A. O., Kallmes, D. F., Kinsman, K. A., ... & Liu, H. (2019). Natural language processing for the identification of silent brain infarcts from neuroimaging reports. JMIR medical informatics, 7(2), e12109. DOI: 10.2196/12109

[13] Ouyang, X., Zhou, P., Li, C. H., & Liu, L. (2015). Sentiment Analysis Using Convolutional Neural Network. 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing. https://doi.org/10.1109/cit/iucc/dasc/picom.2015.349

[14] Santos, I., Nedjah, N., & Mourelle, L. d. M. (2017). Sentiment Analysis using Convolutional Neural Network with fastText Embeddings. IEE. 978-1-5386-3734-0/17/$31.00.

[15] Umer, M., Imtiaz, Z., Ahmad, M., Nappi, M., Medaglia, C., Choi, G. S., & Mehmood, A. (2022). Impact of convolutional neural network and FastText embedding on text classification. Multimedia Tools and Applications, 1-17. https://doi.org/10.1007/s11042-022-13459-x

[16] Barba, P. (2019). Sentiment Accuracy: Explaining the Baseline and How to Test It. Retrieved from https://www.lexalytics.com/lexablog/sentiment-accuracy-baseline-testing#:~:text=Setting%20a%20baseline%20sentiment%20accuracy,training%20a%20sentiment%20scoring%20system.