# Literature Review

## Introduction

Sentiment analysis is used to identify the emotional tone behind a piece of text, generally to gain an understanding of the opinions and attitudes people have towards a particular topic, such as a new product from a business. However a less researched area of sentiment analysis can be using detection on individual reactions on social media such as detecting potential cyber bullying or signs of depression in users (Ciaburro, Iannace, Puyana-Romero, 2022, Wei, Zou 2019).

Twitter has a very large user base and is often used to express opinions which is very useful for sentiment analysis. The twitter API is also accessible to use to create datasets (Zulfadzli, Haliyana, 2019).

This review will first discuss methods to use to construct a helpful dataset to train and test the classifier on, before discussing different machine learning models and feature classification methods that can be used for sentiment analysis.

## Datasets

When working with machine learning algorithms, datasets are used to train the machine and to test the models used. Even though Twitter is one of the most common places to collect data, there aren’t many public datasets available. A lot of the datasets are small or about specific topics such as healthcare or gas prices (Duong, Truong, 2019). Small datasets produce less accurate results since there is less data for the machine to train on and learn from (Zulfadzli, Haliyana, 2019). The largest twitter dataset publicly available is the Stanford Twitter sentiment corpus that has 1.6 million tweets (Saif et al., 2013), however they were labelled as positive or negative automatically instead of manually so are not accurately classified.

### Data Pre-Processing

Twitter is a social media that has a large amount of noise in its data that a machine will have a difficult time sorting through, so it is important to pre-process the data.

There are three different types of text formats in Natural Language Processing: structured sentiments, semi-structured, and unstructured. Tweets are unstructured sentiments which are the most difficult to work with (Duong, Nguyen, 2021). (Duong, Nguyen, 2021) applied the most common pre-processing techniques to four different datasets and observed that the F-score of every dataset improved afterwards. However they did not analyse which individual techniques were most effective.

URL removal is a common pre-processing technique as URLs are bulky and often do not add insight into sentiment. While (Singh, Kumari, 2016) argue that removing URLs is not necessary as it does not improve the later performance of machine learning algorithms, (Jianqiang, Xiaolin, 2017) show that while removing them might not improve performance it also does not decrease performance the way an important feature would, meaning removing URLs is beneficial as it will improve computational performance and reduce cost. This same method was used to prove that stopwords should be removed as well (Saif et al., 2014).

(Lin, He, 2009) argue that removing numbers reduces classification effectiveness, more recent papers show that numbers are mostly irrelevant and do not help with classification, and they can be removed if any emoticons have been turned into words beforehand (Jianqiang, Xiaolin, 2017).

Both (Duong, Nguyen, 2021) and (Duong, Truong, 2019) observed that replacing negations with a word’s antonym is not an effective technique, and while (Duong, Truong, 2019) analysed this effect from the point of view of news classification they used identical methods to tweet analysis and therefore their conclusions can be applied to tweet pre-processing.

(Singh, Kumari, 2016) remove all repeating letters, turning ‘coooool’ into ‘cool’, while (Jianqiang, Xiaolin, 2017) attempts to maintain the sentiment by conserving one extra letter, turning ‘coooool’ into ‘coool’, however this method had different results on all the datasets it was tried on so is not a conclusively effective pre-processing method.

(Duong, Nguyen, 2021) tested different techniques and found that replacing emoticons with words, removing punctuation, stopwords, and numbers, and elongated word handling are effective pre-processing techniques. Their research is done on a Vietnamese dataset which might differ from an English dataset; however their findings are largely supported by (Jianqiang, Xiaolin, 2017), (Effrosynidis, Symeonidis, Arampatzis, 2017) and (Duong, Truong, 2019)who all tested on English datasets.

(Effrosynidis, Symeonidis, Arampatzis, 2017) created a thorough literature review that found that stemming, replacement of repetitions of punctuation, and removing numbers all had positive effects on the models used. The pre-processing techniques that they observed having negative effects on the model were: removing punctuation, turning all capitalized words into lowercase, replacing slang, replacing negations with antonyms, and spelling corrections.

The literature finds that the pre-processing techniques considered most effective are removing pre-defined stopwords, expanding acronyms, replacing emoticons with text equivalents, removing URLs, removing numbers, and removing repeat letters.

While data pre-processing is overwhelmingly seen as a very important step in NLP, (Saif et al., 2014) find that pre-processing leads to a significant reduction of vocabulary size, 62% in their tests. This is a non-negligeable loss of vocabulary; however this can potentially be helped by data augmentation techniques that will insert new vocabulary into the data.

### Data Augmentation

Data augmentation is the process of artificially adding content to a dataset based on the already existing collected tweets. While (Duong, Nguyen, 2021) state that data augmentation is more accurate on larger datasets as there is more vocabulary to use, (Wei, Zou, 2019) and (Duong, Truong, 2019) support the fact that the point of data augmentation is to use it on smaller datasets and is particularly helpful on unbalanced datasets (Abonizio, Paraiso, Barbon, 2022). Twitter datasets can be unbalanced, (Gaind, Syal, Padgalwar, 2019) states that most tweets are positive, however they used a small sample size and only collected tweets in India.

One of the most frequently used data augmentation methods is the Easy Data Augmentation (EDA) method. This method consists of four techniques that can be applied individually or together to a dataset: random insertion, synonym replacement, random swap, and random deletion.

(Wei, Zou, 2019), and (Duong, Nguyen, 2021) use this method in their research to analyse the effect and note that using EDA improves performance. (Wei, Zou, 2019), also noted that when using 50% of the available data in the dataset paired with EDA provided the same performance as when using 100% of the data without DA, meaning that EDA is very effective in augmenting data. While (Wei, Zou, 2019) did not look specifically at sentiment analysis they did text classification that can be applied to sentiment analysis. However, since longer sentences can absorb more changes without losing their meaning, there is a possibility this would not be as effective on tweets where the text is very short.

Back translation is another popular data augmentation method, where a tweet is translated into another language and then back again to create a different sentence with the same meaning. (Yu et al., 2018) showed this method was effective, however it uses more computation effort for close to the same improvement as using EDA.

Type swap can also be used, this method replaces words in the text by other words of the same type (Raiman, Miller, 2017). (Yu et al., 2018) argue that text samples will it be diverse enough as the sentence structures do not change with this method, but it improves the accuracy.

## Algorithms

There are two approaches to classifying tweets: lexicon-based or machine learning.

### Lexicon based algorithms

Lexicon based approaches do not use datasets, they use lexicon dictionaries to count the number of positive and negative words in a text and base the classification on it. This makes them a very easy to implement and cost and energy efficient solution. However (Hamdan, Bellot, Bechet, 2015) commented that lexicon-based approaches weren’t optimal since the performance depends on how good the dictionary is, and that it cannot classify nuances in languages accurately such as sarcasm and negation. Their literature review was conducted in 2015, however (Zulfadzli, Haliyana, 2019) also support this analysis. Lexicon based approaches are particularly unsuitable on their own for twitter sentiment analysis since twitter has a large amount of slang that will not be able to be captured by the dictionary unless it is kept up to date regularly.

(Venkateswarlu, Nandhini, Naulegari, 2019) classify reviews and argue that lexicon-based approaches are better since machine learning classifiers are trained on limited data and may struggle to classify outside of the dataset that it was trained on, however previous research has established that a larger dataset can minimize that issue, and lexicon-based classifiers are also trained on limited data due to the limited capacity of a lexicon dictionary.

Lexicon approaches can still be useful for hybrid classification methods, (Mody et al., 2018) try to identify tweets using lexicon-based approaches to classify the tweets based on emoticons present, and then using machine learning methods on these pre-classified tweets. This method gave good results and shows that using hybrid methods can be effective for tweet sentiment analysis.

### Machine learning Algorithms

Machine learning is different from lexicon approaches as it uses datasets to identify the sentiment of the tweets. Tweets must first be turned into a machine-readable format by turning them into numerical vectors called vectorization.

#### Vectorization

The two main vectorization models are TF-IDF and Bag of Words. While Bag of Words simply counts the amount of time a word is present and adds this to a vector, TF-IDF also considers how often the word is used in other tweets to give a weight to each word. (Abubakar, Umar, Bakale, 2022) finds that TF-IDF perform better than the Bag of Words approach, however they performed this study on book reviews which might differ from tweets as reviews will usually be much longer. (Naf’an et al., 2019) used TF-IDF to flag potential cyber bullying tweets using a strong methodology and found that TF-IDF was an effective vectorizing method, however they did not compare it to Bag-of-Words. TF-IDF is good for more nuanced rankings of sentiments that have more classes than just ‘positive’, ‘negative’, and ‘neutral’, as (Naf’an et al., 2019) does.

TF-IDF generally performs much better than Bag of Words vectorization, (Rakhmanov, 2020) used a 5-class classification and several different machine learning algorithms on a very large dataset and TF-IDF performed better on all of them. While this study was focused on student comments rather than tweets the sentiment analysis can be applied to tweets, however there might be differences again concerning the length of the text used.

(Ismail, Harous, Belkhouche, 2016) looked at whether Bag of Words or TF-IDF was better for twitter sentiment analysis, and they found an inconclusive result. However they performed these tests on a small dataset of only 369 tweets and using data from 2009, since tweets are now longer than they were back then this might not give accurate results for our purpose.

#### Machine Learning Approach

There are several machine learning approaches that come up the most often in the literature: Naïve Bayes, Logistic Regression, Support Vector Machine, and Maximum Entropy. Literature disagrees on which one is best for classification tasks with many papers finding different results. (Zulfadzli, Haliyana, 2019) performed a literature review of papers up to 2019 that found that SVM and Naïve Bayes, which are the most common models for sentiment analysis, perform very similarly to each other. Naïve Bayes is more successful on well formed data which would indicate that it is not ideal for tweets that often have slang and poor spelling or structure, however this could be improved with good pre-processing.

When comparing SVM and Naïve Bayes on airline reviews, SVM performs much better than Naïve Bayes, (Rahat, Kahir, Masum, 2019) test this on tweets and used good pre-processing methods, however the datasets were mostly negative tweets which could skew the results. Naïve Bayes and SVM both have good precision among other machine learning algorithms (Alasaeedi, Khan, 2019), however they recommend hybrid classification as it seems to have the best performance.

### Hybrid Methods

While simple machine learning models are not proven to be necessarily more efficient than others, literature is clear that using hybrid approaches with both machine learning algorithm and a lexicon algorithm produces the best results. This is shown by (Alasaeedi, Khan, 2019) and (Hamdan, Bellot, Bechet, 2015) who test with and without a hybrid approach and consistently find that hybrid approaches perform best.

While (Dhaoui, Webster, Tan, 2017) did not find that hybrid methods were necessarily more efficient, they used a small Facebook dataset and used two different algorithms to classify positive and negative comments separately, making it difficult to reach an accurate conclusion about their performance.

## Conclusion

Research has been done to classify tweets into positive or negative sentiments, however not much research has been made into using this method to detect tweets based on a sliding scale of positivity/negativity instead of a binary approach, even though this could be helpful for cyber bullying detection among others. There are also few datasets to use for sentiment analysis.

# References

SAIF, H. et al., 2013. *Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold.* Turin, Italy: The Open University

ZULFADZLI D., HALIYANA K., 2019. Sentiment Analysis in Social Media and Its Application: Systematic Literature Review. *Procedia Computer Science*. [online]. In Press. Available from: <https://www.sciencedirect.com/science/article/pii/S187705091931885X> [Accessed 20/10/2022]

DUONG, HT., NGUYEN-THI, TA., 2021. A review: preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*. [online]. In Press. Available from: <https://computationalsocialnetworks.springeropen.com/articles/10.1186/s40649-020-00080-x#citeas> [Accessed 13/10/2022]

DUONG, HT., TRUONG HOANG, V., 2019, A Survey on the Multiple Classifier for New Benchmark Dataset of Vietnamese News Classification. *11th International Conference on Knowledge and Smart Technology (KST)*. 23-26 January 2019. Phuket, Thailand: IEEE. p 23-28

SINGH, T., KUMARI M., 2016. Role of Text-Pre-processing in Twitter Sentiment Analysis. *Procedia Computer Science.* [online]. In Press. Available from: <https://www.sciencedirect.com/science/article/pii/S1877050916311607> [Accessed 15/10/2022]

JIANQIANG Z., and XIAOLIN G., 2017. Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. *IEEE Access*. [online]. In Press. Available from: <https://ieeexplore.ieee.org/document/7862202> [Accessed 15/10/2022]

LIN C., HE Y., 2009. Joint sentiment/topic model for sentiment analysis. *18th ACM conference on Information and knowledge management*. 2-6 November 2009. New York, NY: Association for Computing Machinery p375–384.

SAIF, H., et al., 2014. On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In: *LREC 2014, Ninth International Conference on Language Resources and Evaluation.* May 2014.European Language Resources Association. p. 810–817.

EFFROSYNIDIS, D., SYMEONIDIS, S., ARAMPATZIS, A. 2017. A Comparison of Pre-processing Techniques for Twitter Sentiment Analysis. *Research and Advanced Technology for Digital Libraries*, pp.394–406.

WEI, J., and ZOU, K., 2019*. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks*. [online] Available at: <https://arxiv.org/abs/1901.11196> [Accessed 18/10/2022]

YU, A.W., et al., 2018. *QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension*. [online] Available at: <https://arxiv.org/abs/1804.09541> [Accessed 18/10/2022].

ABONIZIO, H., PARAISO, E., BARBON, S., 2022. Toward Text Data Augmentation for Sentiment Analysis. *IEEE Transactions on Artificial Intelligence.* [online]. In Press. Available from: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9543519> [Accessed 18/10/2022]

GAIND, B., SYAL, V., PADGALWAR, S., 2019. *Emotion Detection and Analysis on Social Media.* [online] Available at: <https://arxiv.org/abs/1901.08458> [Accessed 20/10/2022]

RAIMAN, J., and MILLER, J., 2017. *Globally Normalized Reader*. [online]. Available at: <https://arxiv.org/abs/1709.02828> [Accessed 13/10/2022]

CIABURRO, G., IANNACE, G., PUYANA-ROMERO, V., 2022. Sentiment Analysis-Based Method to Prevent Cyber Bullying. *Proceeding of 2021 International Conference on Wireless Communications, Networking and Applications*. Jilin, China. p.721–735.

VENKATESWARLU B., NANDHINI K., NAULEGARI J., 2019. A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis*. Asian Journal of Computer Science and Technology*. [online]. In press. Available from: <https://www.researchgate.net/publication/333602124_A_Comprehensive_Study_on_Lexicon_Based_Approaches_for_Sentiment_Analysis> [Accessed 20/10/2022]

HAMDAN, H., BELLOT, P., BECHET, F., 2015. Lsislif : Feature Extraction and Label Weighting for Sentiment Analysis. *Proceedings of the 9th international workshop on semantic evaluation.* 2015. USA: The Association for Computational Linguistics. p. 568

MODY, A., PIMPLE, R., SHAH, S., SHEKOKAR, N., 2018. Identification of Potential Cyber Bullying Tweets using Hybrid Approach in Sentiment Analysis. *Third International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT).* December 2018. New Jersey: IEEE.

ABUBAKAR, H.D., UMAR, M., BAKALE, M.A., 2022. Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec. *SLU Journal of Science and Technology.* July 2022. p.27–33.

NAF’AN, M.Z., BIMANTARA, A.A., LARASATI, A., RISONDANG, E.M. and NUGRAHA, N.A.S., 2019. Sentiment Analysis of Cyberbullying on Instagram User Comments. *Journal of Data Science and Its Applications*, Indonesia: Informatic Department, Institut Teknologi Telkom Purwokerto Jalan D.I. Panjaitan. p.88–98.

RAKHMANOV, O., 2020. A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify Student-Lecturer Comments. *Procedia Computer Science*. [online]. In Press. Available from: <https://www.sciencedirect.com/science/article/pii/S1877050920323954> [Accessed 20/10/2022]

ISMAIL, H., HAROUS, S., BELKHOUCHE, B., 2016, *A Comparative Analysis of Machine Learning Classifiers for Twitter Sentiment Analysis.* Al Ain, UAE: United Arab Emirates University

RAHAT, A., KAHIR, A., MASUM, A., 2019. Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset. *8th International Conference System Modeling and Advancement in Research Trends (SMART).* 22-23 November 2019. Moradabad, India: IEEE. pp. 266-270.

ALSAEEDI, A., KHAN, M., 2019. A Study on Sentiment Analysis Techniques of Twitter Data. *International Journal of Advanced Computer Science and Applications.* [online]. In press. Available from: <https://www.researchgate.net/publication/331411860_A_Study_on_Sentiment_Analysis_Techniques_of_Twitter_Data> [Accessed 27/10/2022]

DHAOUI, C., WEBSTER, C., TAN, L., 2017. Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing.* [online]. In press. Available from: <https://www.emerald.com/insight/content/doi/10.1108/JCM-03-2017-2141/full/html> [Accessed 29/10/2022]

TUSAR, T., ISLAM, T., 2021. A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data. *2021 International Conference on Electronics, Communications and Information Technology (ICECIT),* 14-16 September 2021. Khulna, Bangladesh: IEEE. p. 1-4