CHAPTER 2

LITERARURE REVIEW

2.1 Sentiment Analysis

Based on Cambridge Dictionary, sentiment is a thought, opinion or idea based on feeling about a situation or a way of thinking about something. While analysis is the act of studying or examining in detail, in order to discover or understand more about it, or your opinion and judgement after doing this. Sentiment Analysis is a task of Natural Language Processing (NLP) that aims to extract sentiments and opinions from texts.(Birjali et al., 2021) Opinion and sentiment can be classified into three main types which are regular opinions, comparative opinion and suggestive opinions. Regular opinions refer to a single entity, comparative opinions compare or contrast more than one entity and suggestive opinions suggest a single or multiple entities. (Shayaa et al., 2018)

In the age of modern science, everything is based on online and on the internet. With the exponential growth of social media, the availability of public opinions and sentiments has increased, making sentiment analysis a crucial tool for comprehending public sentiment across various domains such as business, politics, and others. (Tan et al., 2023) For example, when tourists want to choose a comfortable hotel for their trip, they will look for reviews from other travellers. The reviews in the internet are more trusted than in the brochure because the user’s review can be verified. Those opinions and sentiments are very relevant to our daily lives, and hence there is a need to analyze this user-generated data in order to automatically monitor the public opinion and assist decision-making. (Birjali et al., 2021)

For this reason, the field of sentiment analysis gained more interest within the last one and a half decades among research communities. Since 2004, sentiment analysis has become the fastest growing and the most active research area, as there has been a massive increase in the number of papers focusing on sentiment analysis recently. (Mäntylä et al., 2018) Figure 1.1 shows the rising popularity of sentiment analysis according to Google Trends.

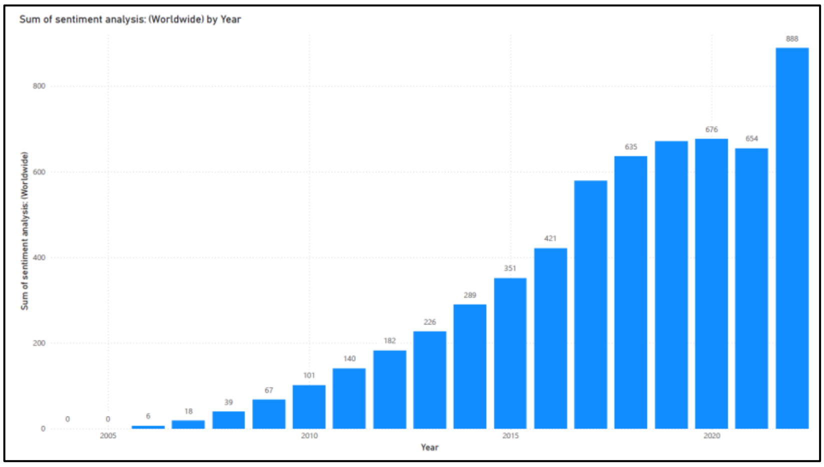

Figure 1.1 Trends of sentiment analysis according to Google Trends(trends.google.com/trends)

2.2 Sentiment Analysis Levels

Sentiment Analysis has been investigated on several levels which are Document Level, Sentence Level, Phrase Level and Aspect Level. Figure 1.2 shows sentiment analysis in each level such as document, sentence, phrase and aspect. Document level sentiment analysis is performed on a whole document and single polarity is given to the whole documents. It provides a single sentiment label for the document as a whole such as positive, negative or neutral. At this level, both supervised and unsupervised learning approaches can be utilized to classify the document. (Bhatia et al., 2015)

Sentence level of sentiment analysis is analyse each sentence and find with corresponding polarity. It focuses on determining the sentiment by individual sentences to positive, negative or neutral. This is highly useful when a document has a wide range and mix of sentiments associated with it. (Yang & Cardie, 2014) Each sentence polarity will be determined independently using the same methodologies as the document level but with greater training data and processing resources. (Wankhade et al., 2022) For example, consider a hotel review "The battery life is excellent. However, the food is not good". So, it will analyse sentence by sentence. “The battery life is excellent” is positive and “However, the food is not good” is negative.

Phrase level of sentiment analysis also be performed where opinion words are mined at phrase level and classification will be done. (Wankhade et al., 2022) This may be useful product reviews of multiple lines; here, it is observed that a single aspect is expressed in a phrase (Thet et al. 2010). It goes deeper than sentence-level sentiment analysis by examining smaller linguistic units, such as clauses or phrases, and labelling their sentiment as positive, negative or neutral. For example, “The food was excellent but the service was terrible”. It will analyse by phrase which “The food was excellent” is positive “but the service was terrible" is negative. Additionally, the term chosen for expression represents the demographic characteristics of individuals, such as gender and age, and its desire, social standing, and personality, other psychological and social characteristics. (Flek, 2020)

Lastly, aspect level is where sentiment analysis is performed. It focuses on identifying sentiment associated with specific aspects or attributes of an entity, rather than analyzing the sentiment of an entire document, sentence, or phrase. Primary attention to all the aspects used in the sentence and assigns polarity to all the aspects after which an aggregate sentiment has calculated for the whole sentence. (Schouten & Frasincar, 2015) For example, the sentence is “The display is beautiful, but the sound quality is disappointing, and the battery life is average.”, the aspect extraction is display, sound quality and battery life. The sentiment classification for display is positive, sound quality is negative and battery life is neutral.

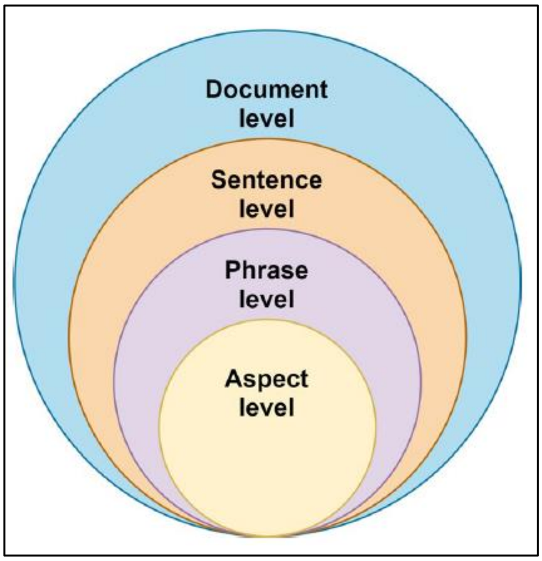

Figure 1.2 Sentiment analysis level (Wankhade et al., 2022)

2.3 Sentiment Analysis Approach

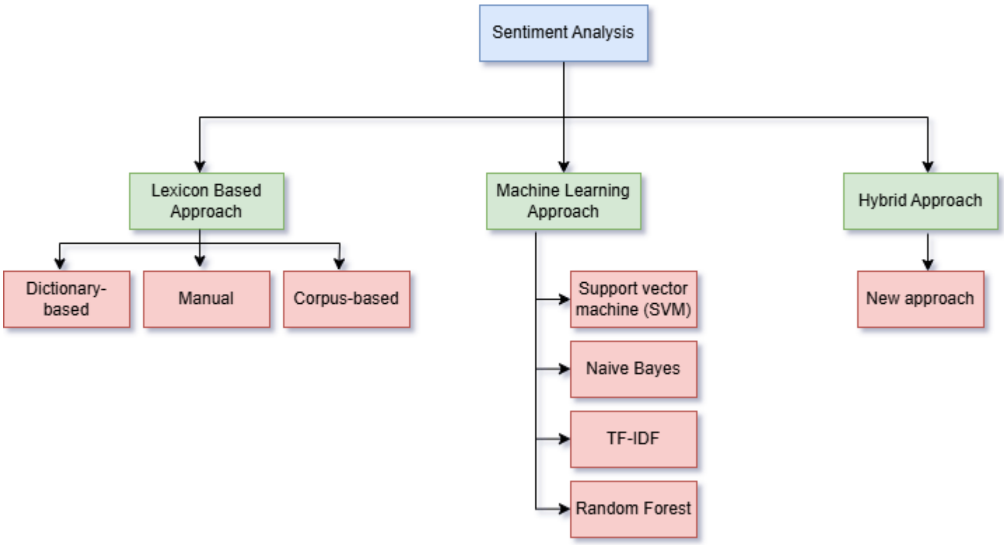

Figure 1.3 Sentiment analysis approach

Figure 1.3 shows the sentiment analysis approach. Sentiment analysis has a few approaches which are Machine Learning, Lexicon-based and Hybrid. Machine learning based approach uses classification technique to classify text. It consists of two sets of documents which are training set and a test set. The training set is used for learning the differentiating characteristics of a document, while the test set is used for checking how well the classifier performs. (Alessia et al., 2015) Support Vector Machine (SVM) is one of the most famous supervised machine learning based classification algorithm. (Khairnar & Kinikar, 2013). Naïve Bayes (NB) is a simple classifier and it is one of the most commonly used algorithms in the field of text classification. (Birjali et al., 2021) Random Forest targets the enhancing and storing of classification trees. (Breiman, 2001), and Term Frequency–Inverse Document Frequency (TF-IDF).

Lexicon-based approach the collection of tokens where each token is assigned with a predefined score which indicates the neutral, positive and negative nature of the text (Kiritchenko et al., 2014). Based on (Wankhade et al., 2022), there are three techniques to construct a sentiment lexicon which are manual construction, corpus-based methods and dictionary-based methods. The manual construction is a difficult and time-consuming task. Corpus-based methods can produce opinion words with relatively high accuracy. Finally, in the dictionary-based techniques, the idea is to first collect a small set of opinion words manually with known orientations, and then to grow this set by searching in the WordNet dictionary for their synonyms and antonyms.

Lastly, Hybrid approach is the combination of machine learning and Lexicon-based approach. The main reason behind the hybrid approach is to inherit high accuracy from machine learning and stability from lexicon-based approach. (Birjali et al., 2021) The focus of this research is more into machine learning. For example, (Amrani et al., 2018) proposed using machine learning-based hybrid approach including RF and SVM. They have shown that the individual models of SVM and RF had an accuracy of 81.01 and 82.03 percent, respectively, whereas the hybrid model combining both the algorithms had an accuracy of close to 84% in the product review dataset provided by amazon.com. Table 2.1 shows the possible advantages and limitation of sentiment analysis approach.

Type of approach	Advantages	Limitation
Machine Learning	The ability to adapt and create trained models for	The low applicability to new data because it is necessary the availability of labelled

	specific purposes and contexts	data that could be costly or even prohibitive
Lexicon-based	Wider term coverage, labelled data and the procedure of learning is not required.	Finite number of words in the lexicons and the assignation of a fixed sentiment orientation and score to words
Hybrid	Lexicon/learning symbiosis, the detection and measurement of sentiment at the concept level and the lesser sensitivity to changes in topic domain	Noisy reviews

Table 2.1 The Advantages and Limitation of Sentiment Analysis Approach

2.4 Machine Learning Approach

Machine learning works better for sentiment analysis but creating an accurate machine learning models is not easy. (Khomsah, 2020) There are many ways to optimize machine learning models, including techniques, evolutionary machine learning algorithms, and intelligent swarms. (Rizaldy & Santoso, 2017) The machine learning algorithm TF-IDF, Random Forest, TF-IDF and Naïve Bayes will be discussed.

2.4.1 The Term Frequency-Inverse Document Frequency (TF-IDF)

The term frequency-inverse document frequency (also called TF-IDF), is a well-recognized method to evaluate the importance of a word in a document. *TF* which counts the number of times a term word appears in the document Because each document is varied in length, it is likely that a term will appear far more frequently in longer documents than in shorter ones. (Wankhade et al., 2022)

$$Term\ Frequency = \frac{(Number\ of\ times\ term\ t\ present\ in\ a\ document)}{(Total\ number\ of\ terms\ in\ the\ document)}$$

IDF (Inverse Document Frequency) is used to give lower weight to words that occur frequently and to give larger words to words that occur rarely. (Qaiser & Ali, 2018) There are some terms like “is”, “an”, “and”, “when” and others which occurs frequently but do not have any importance. IDF is calculated as IDF (t) = log(N/DF), where N is the number of documents and DF is the number of document containing term t.(Ahuja et al., 2019) Suppose there is a document which contains 200 words and out of these 200 words mouse appears 10 times than term frequency will be 10/250=0.04 and suppose there are 50000 documents and out of these only 500 documents contains mouse. Then, IDF (mouse) = 50000/500=100, and TF-IDF (mouse) will be 0.04*100= 4.

2.4.2 Random Forest

Random forest is an ensemble method which creates a number of decision trees turned into an ensembled forest. (Zahid-samza595, 2020) Each tree in the Random Forest returns a class prediction and Random Forest makes decision based on the majority of votes. (Saad & Aref, 2020) As defined random forest is a classifier comprised of tree structured classifiers {h(x,Ok), k=1,…} having {Ok} as independently identically distributed random vectors and each tree casts a unit vote for the most popular and famous class at input x. (Ahmad et al., 2017) Each tree is fabricated by using a bootstrap sample of the data, and the prediction of all trees are ultimately accumulated by means of majority voting. (Shayaa et al., 2018)

2.4.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the most effective and famous classification machine learning methods. (Saad & Aref, 2020) SVM are a type of non-probabilistic supervised learning technique that is frequently used for classification tasks. SVM primary objective is to determine the hyperplane that best separates the data into distinct classes. (Wankhade et al., 2022) Text classification using SVM is done by representing each document as a vector where dimensions is the number of unique keywords used in data used for training.

(Zahid-samza595, 2020) SVM is applicable for the data which are linearly separable and also it applicable non-linear data with proper kernel functions or tricks. (Bania, 2020) Figure 1.4 shows the SVM illustration the left is the original objects and the right is mapped or rearranged using a mathematical function known as kernel and this is known as mapping or transformation. After transformation, the mapped objects are linearly separable and as a result the complex structures having curves to separate the objects can be avoided. (Devika et al., 2016)

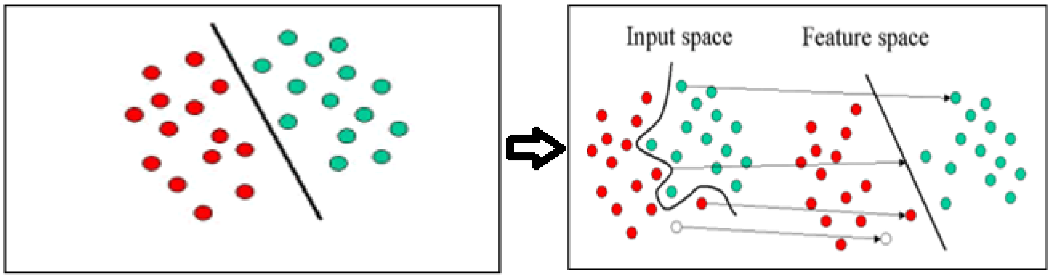

Figure 1.4 SVM illustration (Devika et al., 2016)

2.4.4 Naïve Bayes (NB)

Naïve Bayes (NB) is a powerful algorithm for classification used for classifying data on basis of probabilities. (Ahuja et al., 2019) In machine learning it is in family of sample probabilistic classifier based on Bayes theorem and depends on BoW feature extraction.. (Devika et al., 2016) It normally entails a minimal data set for training and then is used to predict the parameters needed for classification purposes. (Shayaa et al., 2018) It works efficiently with large datasets as it is fast and accurate with a very low computational cost. (Saad & Aref, 2020) Figure 1.5 shows the Bayes Theorem used in NB.

$$\boxed{P(c|D) = \frac{P(D|c) * P(c)}{P(D)}}$$

Figure 1.5 Bayes Theorem (Zahid-samza595, 2020)

2.5 Review of Similar Works

Table 1.1 shows similar works that had been done by previous researchers about the project. All of the researchers use the same features which is the term frequency-inverse document frequency (TF-IDF). The different is the classifier. Makhmudah et al and Alzyout et al use the same classifier which is Support Vector Machine. The accuracy for dataset used by Makhmudah et al is the highest which is 99.5% while Alzyout et al used self-collected dataset has the accuracy 78.25%. Alsalman that used Multinomial Naïve Bayes has the second highest accuracy which is 87.5%. While Rathi et al used AdaBoost which unsupervised method, AdaBoost has the lowest accuracy which is 67%.

Author	Dataset	Features	Classifier	Accuracy (%)
(Rathi et al., 2018)	Sentiment140, Polarity Dataset, and University of Michigan dataset	TF-IDF	AdaBoost	67
(Makhmudah et al., 2019)	Tweets related to homosexuals	TF-IDF	Support Vector Machine (SVM)	99.5
(Gupta et al., 2019)	Sentiment140	TF-IDF	Neural Network	80
(Alsalman, 2020)	Arabic Tweets	TF-IDF	Multinomial Naïve Bayes	87.5
(Alzyout et al., 2021)	Self-collected dataset	TF-IDF	Support Vector Machine (SVM)	78.25

Table 2.2 Example of similar works

2.6 Research Gap

Firstly, the research not enough attention to the other method than Machine Learning. For example, Lexicon-based approach. There are few Lexicon-based approach techniques that widely used to do the classification. This is because the course is more focus on the Machine Learning. Apart from that, the research also not enough attention on the project that the dataset is in other languages. For example, in this project sentiment analysis on hotel reviews. So, the reviews may be in Arabic or Germany language. But, for this research only focus more on the English and eliminate anything other that English language.

REFERENCES

Ahmad, M., Aftab, S., Muhammad, S. S., & ... (2017). Machine learning techniques for sentiment analysis: A review. In *Int. J. Multidiscip. Sci …*. researchgate.net. https://www.researchgate.net/profile/Shabib-Aftab-2/publication/317284281_Machine_Learning_Techniques_for_Sentiment_Analysis_A_Review/links/59302d6ba6fdcc89e78431ec/Machine-Learning-Techniques-for-Sentiment-Analysis-A-Review.pdf

Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. *Procedia Computer Science*. https://www.sciencedirect.com/science/article/pii/S1877050919306593

Alessia, D., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. In *International Journal of Computer …*. Citeseer. https://citeseerx.ist.psu.edu/document?repid=rep1\&type=pdf\&doi=d709747c589bca62d9ee85752fb3a8ec899cac20

Amrani, Y. A., Lazaar, M., & Kadiri, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science*. https://www.sciencedirect.com/science/article/pii/S1877050918301625

Bania, R. K. (2020). COVID-19 public tweets sentiment analysis using TF-IDF and inductive learning models. *INFOCOMP Journal of Computer Science*. https://infocomp.dcc.ufla.br/index.php/infocomp/article/view/985

Bhatia, P., Ji, Y., & Eisenstein, J. (2015). Better document-level sentiment analysis from rst discourse parsing. *arXiv Preprint arXiv:1509.01599*. https://arxiv.org/abs/1509.01599

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*. https://www.sciencedirect.com/science/article/pii/S095070512100397X

Breiman, L. (2001). Random forests. *Machine Learning*. https://doi.org/10.1023/a:1010933404324

Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: A comparative study on different approaches. *Procedia Computer Science*. https://www.sciencedirect.com/science/article/pii/S187705091630463X

Flek, L. (2020). Returning the N to NLP: Towards contextually personalized classification models. *Proceedings of the 58th Annual Meeting of the …*. https://aclanthology.org/2020.acl-main.700/

Khairnar, J., & Kinikar, M. (2013). Machine learning algorithms for opinion mining and sentiment classification. In *International Journal of Scientific and Research …*. Citeseer. https://citeseerx.ist.psu.edu/document?repid=rep1\&type=pdf\&doi=269d91e79049092bdf0651241d0d66830aa9fafc

Khomsah, S. (2020). Naive bayes classifier optimization on sentiment analysis of hotel reviews. *Jurnal Penelitian Pos Dan Informatika*, *Query date: 2024-12-01 16:20:27*. https://jurnal-ppi.kominfo.go.id/index.php/jppi/article/view/322

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*. https://www.sciencedirect.com/science/article/pii/S1574013717300606

Qaiser, S., & Ali, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to documents. In *International Journal of Computer Applications*. researchgate.net. https://www.researchgate.net/profile/Shahzad-

Qaiser/publication/326425709_Text_Mining_Use_of_TF-IDF_to_Examine_the_Relevance_of_Words_to_Documents/links/5b4cd57fa6fdcc8dae245aa3/Text-Mining-Use-of-TF-IDF-to-Examine-the-Relevance-of-Words-to-Documents.pdf

Rizaldy, A., & Santoso, H. A. (2017). Performance improvement of Support Vector Machine (SVM) With information gain on categorization of Indonesian news documents. *2017 International Seminar on …*. https://ieeexplore.ieee.org/abstract/document/8251874/

Saad, S., & Aref, M. (2020). A survey on sentiment analysis in tourism. *... Journal of Intelligent Computing and Information …*. https://journals.ekb.eg/article_106309.html

Schouten, K., & Frasincar, F. (2015). Survey on aspect-level sentiment analysis. *... Transactions on Knowledge and Data …*. https://ieeexplore.ieee.org/abstract/document/7286808/

Shayaa, S., Jaafar, N. I., Bahri, S., Sulaiman, A., Wai, P. S., & ... (2018). Sentiment analysis of big data: Methods, applications, and open challenges. *Ieee …*. https://ieeexplore.ieee.org/abstract/document/8399738/

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A survey of sentiment analysis: Approaches, datasets, and future research. In *Applied Sciences*. mdpi.com. https://www.mdpi.com/2076-3417/13/7/4550

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. In *Artificial Intelligence Review*. Springer. https://doi.org/10.1007/S10462-022-10144-1

Yang, B., & Cardie, C. (2014). Context-aware learning for sentence-level sentiment analysis with posterior regularization. In *Proceedings of the 52nd Annual Meeting of the …*. aclanthology.org. https://aclanthology.org/P14-1031.pdf

Zahid-samza595, S. (2020). *Sentiment analysis of hotel reviews-performance evaluation of machine learning algorithms*. researchgate.net. https://www.researchgate.net/profile/Saman-Zahid/publication/351262735_Sentiment_Analysis_of_Hotel_Reviews_-_Performance_Evaluation_of_Machine_Learning_Algorithms/links/608db79b92851c490fae3a1e/Sentiment-Analysis-of-Hotel-Reviews-Performance-Evaluation-of-Machine-Learning-Algorithms.pdf