# Text Mining

Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights. Text mining is an artificial intelligence (AI) technology that uses natural language processing (NLP) to transform the free (unstructured) text in documents and databases into normalized, structured data suitable for analysis or to drive machine learning (ML) algorithms.

# Sentiment Analysis

**Sentiment analysis** (also known as **opinion mining** or **emotion AI**) is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

# RQ1- The sentiment analysis method used in social media

Based on the papers reviewed, all of the paper demonstrated the usage of either Lexicon based method, Machine learning method or a mix of both method when implementing sentiment analysis.

Lexicon based method is known as an unsupervised learning method. Lexicon method does not require any training data and only depends on the dictionary. Most of the study adapted Sentiwordnet and TF-IDR method when conducting sentiment analysis. This approach is calculated based on the occurrences of the terms in the text data with other positive or negative words in the predeveloped polarity lexicons like Sentiwordnet. As for TF-IDR method, it works by converting the words into a number and it is calculated using the term frequency-inverse document frequency method.

Due to the complexity of natural languages, this approach is not designed to cover all aspect of language especially when it comes to slang, sarcasm, and negation. Using just sentiment words are not sufficient.

The lexicon-based method does have its own advantage such as it provides simple counting positive and negative words, flexible to fit with different language and speed to complete analysis.

Machine learning method falls under supervised learning and the method requires training data in order to be processed. The most used method in machine learning method is the SVM and Naïve Bayes model. Naïve Bayes is successful when applied on well-formed text corpus while support vector machine (SVM) gives a good performance for low shape dataset. Also, analyzing with machine learning is time consuming where it takes hours in the complex machine learning model especially if training is required. The process is faster with a smaller size of training dataset, but it leads to poorer classification accuracy.

In order to improve the outcome, it is recommended to combine both methods as it will complement each other, and the result is improved compared to using one approach only.

# RQ2 - Type of social media platform use to extract data for sentiment analysis

Social information services or social media can be categorized into four types based on their application usage: Content communities (YouTube, Instagram), Social networking (Facebook, LinkedIn), Blogs (Reddit, Quora) and Micro-blogs (Twitter, Tumblr).

Based on the reviewed paper, among the four types of social media services, micro-blogging sites specifically twitter is the top social media platform used to collect information on user opinion

Twitter as a famous microblogging tool social media platform for people to express their emotion towards a particular person, event, or product. What makes Twitter popular is the content or data that is readily available for public use. With the usage of API, people can access and copy the data on any desired topic based on the keywords or hashtag.

Facebook has the largest social media users in the world. But it is not very popular for sentiment analysis as the data is messy, it is not structured well, and people often use short forms and a lot of spelling error. This makes the data harder to be analyzed.

The other sources of social media are not preferable because of the number of data or opinions that can be extracted is limited such as in Blogspot, YouTube and WordPress.

# RQ3- Application context of sentiment analysis

The application of sentiment analysis ranges from business and marketing, politic, health to public action. Sentiment analysis is not limited to one application, but it provides a vast application in different areas to assist in decision making.

* Sentiment analysis can be applied on world events such as an event, activity, sports, or disaster that is occurring in the world. Sentiment analysis also allows raising awareness of data security and the danger of security breaches. It also acts as a guideline for companies to respond to security breaches in shaping public perception.
* We can see the application of sentiment analysis in healthcare and where the study uses Sentiment analysis as a service framework is proposed and utilize spatio-temporal properties to identify locations of disease outbreaks. In addition, sentiment analysis can identify sentiment needs of people during a disaster and prepare an appropriate response to rescue. Moreover, Sentiment analysis allows finding the level of depression of a person by overserving and analyzing emotions from text.
* Sentiment analysis can be used to predict political election where it shows the data analyzed from twitter is more reliable as a platform where 94% of correlation has been found to polling data and have the potential to become a platform that is able to rival sophisticated polling techniques.
* Feedback of customer plays an utmost important role in the application of sentiment analysis where it can assist business and organization to take appropriate action to improve their product or services and business strategy. Sentiment analysis creates advantages for business owners to evaluate their business flow of stock price through social media.

Top of Form