**Sentiment Analysis for Social Media**

Student No.: L2200191 Name: Timbi Zumnan Bung 祖马 Email: johnsonzeuz@gmail.com

**Abstract**

Opinion mining, also referred to as sentiment analysis, focuses on opinions that separate out good or negative thoughts. Text mining and the NLP model are used in sentiment analysis to extract sentiment (a belief or statement that is held or expressed) from unstructured text data. Sentiment research can be used in industries including social media, customer support, trade, and brand awareness. Sentiment analysis has become the focus of social media research as a result of the development of text mining and natural language processing (NLP) in response to the proliferation of global data throughout time. Sentiment analysis has also experienced a rapid expansion in social media. As social media platforms proliferated, they spawned new features like forum discussions, blogs, reviews, comments, reactions, and postings. These features act as platforms for sentiment analysis and decision-making, and all that is required is the use of the appropriate tools, like the aforementioned NLP, text mining, and machine-learning models like Logistic Regression. In order to extract subjective information from natural languages, sentiment analysis is used. Finally, various techniques will be employed to analyse Twitter data sets. Positive and negative Frequencies (LR Model accuracy 77.92%) and Bag-of-words (LR Model accuracy 88.52%) appear to be less accurate than TF-IDF (LR Model accuracy 88.65%).

***Keywords*:** Sentiment, Opinion, NLP, Twitter

1. **Introduction**

User views have an impact on daily decisions, which can affect many different parts of our daily lives. These decisions might involve everything from picking an institution to spending money on things like gadgets. Prior to the development of the internet, individuals relied on the opinions and reviews of actual users, shoppers, customers, and friends. The development of the internet has made it simpler to assemble several viewpoints from various users and people around the globe. For information on how a specific service or product may be evaluated online or in the marketplace, people can now look for reviews on intended products using websites like CNET, Epinions.com, and e-commerce sites like Taobao, AliExpress, and eBay. They can also use online opinion sites like Yelp and social media platforms like Twitter, Weibo, and Facebook. Additionally, businesses use social media platforms, surveys, and polls to gather feedback on their goods and services. The computational examination of sentiments, emotions, and views expressed in a structured or unstructured text is known as sentiment analysis or opinion mining. Because sentiment analysis produces well-processed information, it makes it easier to monetize services and goods. Customers' feedback on a marketing platform, for instance, can be used to gauge success or failure as well as ranking among others, which can help in altering for a more successful conclusion. [1]

1. **Section 2 Sentiment Analysis**

Sentiment as a word can be defined as a view, attitude, or thought mostly based on emotion instead of a reason.

In order to automatically extract sentiment from an unstructured data text, sentiment analysis, also known as opinion mining, employs Natural Language Processing (NLP), computational techniques like text mining, text normalization, and text vectorization, as well as supervised learning models like logistic regression and unsupervised learning models like lexicon-based, sentiwordnet model, and TF-IDR model. Typically, sentiment analysis just takes into account positive and negative attitudes in order to be clear.[2]

**2.1 Application of Sentiment Analysis in social media**

* Finding like-minded individuals or communities on platforms like Twitter and Weibo
* Tracking Customer and interest Trends on Twitter and Weibo [3]

**2.2 Social media platforms that extract data for sentiment analysis**

Social media applications such as Facebook, LinkedIn

Microblogging platforms such as Weibo and Twitter

Content communities’ platforms such as Bilibili, YouTube, and Instagram

Blogs such as Quora and Reddit

The most popular platform for gathering user feedback among the aforementioned media services is the microblogging site Twitter. Users can engage and send brief messages on Twitter, the most popular platform for doing so. By doing so, users can express their opinions about people, organizations, events, and even products. To conduct an analysis, it is possible to get data on any desired themes based on hashtags and keywords with the help of API. In contrast, while having the most social media users globally, Facebook is not a good candidate for sentiment analysis due to the volume of noise in the data, the unstructured nature of the platform, and the prevalence of slang, shorthand, and spelling mistakes among users.[3]

**2.3 Different levels of sentiment analysis**

**Document Level**

At this level, sentiment may be retrieved from a review as a whole, and opinions are categorized according to the sentiment of the opinion holder as a whole. A review is classified as either favourable, negative, or neutral at the document level. A sample document level is shown below.

“I bought a surface book pro last week. It is such a nice laptop, although a bit small. The touchscreen has good touch responsiveness. The audio quality is standard. I like it!’

It shows that this review is positive. Document level works best when the document is written by an individual who expresses his or her opinion on a single entity.

**Sentence Level**

Sentence level usually involves two steps:

* Subjectivity classification of sentences; Objective and Subjective
* Polarity classification; Positive, Negative, and Neutral

Sentences that are subjective express opinions, sentiments, or personal convictions while sentences that are objective present some factual facts. Subjective sentences can be identified using naive Bayesian categorization. It is not sufficient to just know what people think. Multiple viewpoints, subjective clauses, and factual clauses are all possible with subjectivity. Here is an illustration of a sentence level. “Huawei Phones are doing well despite sanctions by the US government.”

Both Document level and sentence level can’t find people’s likes and dislikes and can’t find opinion targets.

Example of an objective sentence: “Surface book pro is a high-end laptop,”

Example of a subjective sentence: “Surface book pro is awesome,”

Example of a polarity: “I guess the laptop is good”.

**Aspect or Feature Level**

The aspect or feature level seeks to locate and extract object characteristics that have been commented on by an opinion holder and classified as either good, negative, or neutral. Synonyms for features are collected into groups, and a feature-based summary of numerous reviews is created. The aspect or feature-level example is shown below.

“I bought a surface book pro last week. It is such a nice laptop, although a bit small. The touchscreen has good touch responsiveness. The audio quality is standard. It is much better than my previous Surface book 3, which has poor touch responsiveness and poor screen resolution with its large screen. However, my friend was not happy with me because I didn’t tell him before I bought the laptop. He also thought the laptop was too expensive,”

Here we can see Opinion targets (entities and their aspects), Sentiments (positive and negative), Opinion holders (the buyer and his friend), and Time (when the opinion was expressed) [4]

**2.4 How to Explicate Features for Sentiment Detection**

Bag of words: It extracts features from the text so that the text can be used with the machine learning algorithm.

* Annotated lexicons (WordNet, SentiWordNet): An opinion lexicon derived from the WordNet database with numerical scores indicating positive and negative sentiments.
* Syntactic patterns: These rely on predefined Parts of Speech.

The features to be considered in the detection include:

* Words (Unigrams)
* Phrases (n-grams)
* Sentences [5]

1. **Section 3 Definition of Opinion and Emotion**

An opinion can be seen as a quintuple (ej , ajk , soijkl , hi , tl ) [2] 18-54

Where

A target entity is called ej,

and an attribute of that entity is called ajk.

- Soijkl is the sentimental weight of the view expressed by the opinion bearer on a certain aspect of an entity at a given period.

-soijkl represents +ve, -ve, neu, or more precise ratings.

- hi is a person who holds an opinion; - tl is the moment at which that opinion is conveyed; - ej, ajk is sometimes referred to as an opinion target

**3.1 Types of Opinions**

Regular opinions could be sentiment or opinion expression on some target entities.

* Direct opinions: sentiment expression on one object

“The touchscreen has good responsiveness.”

“The audio quality of this laptop is standard”

* Indirect opinions: Differences (objective or subjective) of more than one object

“Surface book pro is more expensive than Surface book 3.” (objective)

“Surface book pro is better than Surface book 3. (subjective)

Comparative opinions: Comparing more than one entity.

“iOS is better than Android.” [2] 18-54

**3.2 Emotion**

Humans communicate their emotions through gestures, postures, and facial expressions, which are ways of expressing their subjective feelings and thoughts. According to a 2017 study that was published in Proceedings of the National Academy of Sciences, there are 27 different types of emotions. The strength of some emotions, such as joy and rage, is mostly related to the sentiment's level. However, this is not always the case with sentiment assessments. Rational and emotional evaluations are the two categories into which evaluations can be separated.

* Emotional Evaluation

“I love MTN,” “I am angry with their customer service,” and “This is the best car ever built.”

* Rational Evaluation

“The camera quality of this phone is good,” “This laptop is worth the price,” “I am happy with this car.” [2] 18-54

1. **Section 4 Sentiment Classification**

An automated procedure called sentiment classification is used to recognize text sentiments and sort them into three categories based on the feelings of the users: positive, negative, or neutral. Sentiment categorization and Natural Language Processing can be used to evaluate subjective data and assist brands and businesses in better understanding how users and consumers feel about their services and goods.[2] 55-58

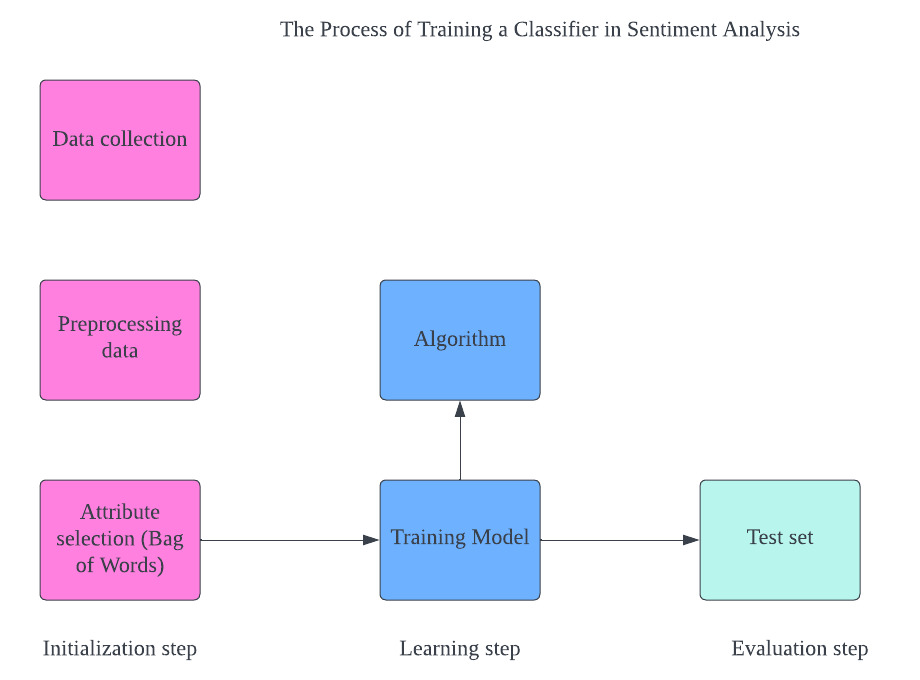


Figure 1. The Process of training a classifier in sentiment analysis.

**4.1 Three Sentiment Analysis Algorithms Models**

Sentiment analysis uses three algorithm models;

* **Rule-Based Systems**

Due to the rule-reliance based on a lexicon of positive, negative, and neutral concepts, rule-based is sometimes referred to as lexicon-based. Without training or employing machine learning models, the rule-based approach provides a useful technique to analyse the text. Lexicon-based systems frequently employ SentiWordNet, TextBlob, and VADER.

Positive sentiment will be returned if the number of positive words in a text is larger than the number of negative words, and vice versa. If both counts of positive and negative words are equal, a neutral sentiment will be returned by the rule-based method.

The rule-based approach has some drawbacks, including the inability to recognize words that are not part of the lexicon and the difficulty it has in separating words from their context in order to discern sarcasm, irony, and polysemy. "Great durability? Far from it!" is an example of a sentence that could be viewed negatively. Scaling and improving the lexicon is challenging because the addition of new terms may also change the lexicon's past results.[6]

The following steps must be considered in order to process data in a rule-based system;

Cleaning text

Tokenization

Part of Speech tagging

Stopwords removal

Obtaining stem words

-***TF-IDF Approach***

It's crucial to have baseline techniques and accuracy levels that may be utilized to gauge gains before putting any Rule-based system methods into practice. Before introducing the baseline method, the TF-IDF score needs to be clarified. The term frequency-inverse document frequency, or TF-IDF, is an acronym. The TF-IDF provides a gauge for the relative weighting of each word within a group of documents.

TF-IDF (t, d, D) = tf (t, d) \* idf (t, d, D)

the- term frequency

idf- inverse document frequency

tf(t, d) = count(t) in document

idf(t, d, D) = -log(P(t/D))

where (P(t/D)) is the likelihood of viewing word t in the event that document D is chosen.

By this time, a vector has been created for each document, and each entry in the vector represents a TF-IDF score. To represent the full set of D, we may put the vectors into a matrix, and we can train a logistic regression classifier—or any other classifier—to predict the overall emotion of D.

***Dictionary-Based Approach***

Each time a new phrase is discovered in datasets, the process is repeated. Meaningful words are gathered manually, and then a list of synonyms and antonyms is created that may later be matched to the list. Synonyms are also sorted.

***Corpus-Based Approach***

The statistical technique and the semantic approach are both applied to a particular subject, whereas the corpus-based approach focuses on addressing languages.

***Statistical approach***

This is used to find the occurrence of words, and to determine positive and negative words. Cosine similarity is a type of statistical approach used to determine polarity and non-zero vectors.

**Automated Systems (Based on Machine Learning)**

Automated systems involve the use of machine learning algorithms. The machine learning models can be divided into two kinds.

**Traditional Models –** In this model, datasets with positive, negative, and neutral terms are gathered, processed, and the algorithm is trained using the terms. The principal application of this approach is to establish the polarity of a text.

Due to their scalability, traditional models incorporate machine learning techniques including Naive Bayes, Support Vector Machine, and Logistic Regression.

We utilize the Logistic Regression to estimate our tweet sentiment because it is scalable.

The process of estimating the association between a dependent variable and one or more explanatory factors is known as logistic regression. For instance, your weight is a dependent variable, and the explanatory variable is the ratio of what you eat to gain weight. In logistic regression, a binary dependent variable is modelled using the logistic function.

Logistic Function

Linear Combination

where are learned parameters

Where x are explanatory input variables

Naïve Bayes is also one of the models in traditional models to be considered for sentiment analysis,

Naïve Bayes model classifiers documents, comments, and text as positive or negative.

**Deep Learning Models** – this model gives more accurate results than the traditional models, this model includes neural network models such as CNN (Convoluted Neural Network), DNN (Deep Neural Networks), and RNN (Recurrent Neural Network). [6]

**Hybrid Systems**

The Hybrid systems are the most advanced, widely-used, and efficient systems for sentiment analysis. A well-designed system is needed to be able to get the extra benefits of both automatic and rule-based systems. It inherits high accuracy and stability from the statistical method and lexicon-based method respectively. [6]

1. **Section 5 Applied Text mining Using Python, and NLP Techniques for Sentiment Analysis for Tweeter Datasets**

The Processes involved are listed below and discussed in full in the HTML extracted code from Jupyter Notebook

* **Text Mining**

Text mining and NLP

* **Text Normalization**

Features cleaning

Tokenization

Stemming

Lemmatization

* **Text Vectorization**

Text Representation

Positive/negative

Bag-of-Words

TF-IDF

* **Sentiment Analysis**

Logistic Regression

Model Training

Model Evaluation

Model prediction



[7]

**Future Directions in Sentiment Analysis for social media**

Unstructured data consumption is on the rise, slang usage in human language is growing more complex, data is sparse, the same words have diverse meanings, and multilingual statements are all challenges that need to be overcome. In order to improve sentiment analysis for social media, we can focus on these issues in future research and come up with better, more effective solutions. To guarantee meaningful changes in outcomes, the Deep learning model might be taken into consideration for further studies.

**Conclusion**

The models such as the Rule-based, Machine learning-based, and hybrid-based were used to experiment and evaluate performances, and sentiment analysis offers accurate emotions underlined in context. Though these outcomes may change when applied to various methodologies, the Rule-based approach, TF-IDF, and Logistic regression of machine learning attained and delivered better sentiment analysis results in our considered tweet datasets.

**References**

1. <https://www.globallogic.com/se/wp-content/uploads/2019/12/Introduction-to-Sentiment-Analysis.pdf>
2. Bing Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions,  
   Cambridge University Press. PP 1-17, 18-54, 55-88. 2015.

1. <https://www.datasciencecentral.com/social-media-sentiment-analysis-using-twitter-datasets/> . 2022.
2. Li, X.; Peng, Q.; Sun, Z.; Chai, L.; Wang, Y. Predicting Social Emotions from Readers’ Perspective. IEEE Trans. Affect. Comput. PP 10, 255–264. 2017.
3. Fouad M.M., Gharib T.F., Mashat A.S. Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble. In: Hassanien A., Tolba M., Elhoseny M., Mostafa M. (eds) The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018). AMLTA 2018. Advances in Intelligent Systems and Computing, vol 723. Springer, Cham. 2018.

1. <https://monkeylearn.com/blog/sentiment-classification/> .2020.

1. <https://towardsdatascience.com/a-beginners-guide-to-sentiment-analysis-in-python-95e354ea84f6> .2020