# Analyze the Emotional Tone in English text using Machine Learning

[1] **Thatha Venkata Nagaraju, [2] Bhavani Govardhan [3] Andhra Rajesh**
[1] Associate Professor, Department of CSE, Rise Krishna Sai Gandhi Group of Institutions,[2] Associate Professor, Department of CSE, Rise Krishna Sai Gandhi Group of Institutions, [3] Assistant Professor, Department of CSE, Rise Krishna Sai Gandhi Group of Institutions.

**Abstract:**
In this paper Emotional means Sentiment of Humans. Sentiment analysis, a subfield of natural language processing (NLP), plays a crucial role in understanding and extracting emotions, opinions, and attitudes expressed in text. This paper focuses on the application of sentiment analysis to English texts, with a specific emphasis on its importance, methodologies, and potential real-world applications. By examining the sentiments conveyed in English language content, we aim to contribute to a better understanding of the emotional landscape in textual data.

**Keywords:** Machine Learning (ML), Natural Language Processing NLP, Sentiment analysis(SA).

## I. Introduction

In our increasingly digital world, the abundance of text data generated through social media, online reviews, news articles, and more has created a rich source of information ripe for analysis. Sentiment analysis, also known as opinion mining, provides a means to harness this wealth of textual information by automating the process of determining the underlying emotional tone. This technique has found applications in areas such as market research, customer feedback analysis, and social media monitoring.

Our research focuses on sentiment analysis applied specifically to English texts, considering the unique nuances and challenges associated with this language. This paper will delve into the methods, tools, and techniques used for sentiment analysis, and it will also explore the significance of this field in today's data-driven decision-making processes. By the end of this paper, readers should gain a comprehensive understanding of how sentiment analysis can be a valuable asset in extracting insights and sentiments from English language content.

Sentiment analysis, also known as opinion mining, is a vital field within natural language processing (NLP) that aims to determine the sentiment or emotional tone expressed in text data. With the exponential growth of online content, social media, and user-generated reviews, sentiment analysis has gained immense significance in understanding public opinion, customer feedback, and market trends. This comprehensive literature review provides an overview of the key developments in sentiment analysis, discusses various techniques and methodologies, and explores future possibilities for enhancing sentiment analysis in the context of machine learning.

Sentiment analysis, also known as opinion mining, is a subfield of natural language processing (NLP) and machine learning that focuses on extracting and understanding subjective information from textual data. It plays a crucial role in various applications, such as social media monitoring, customer feedback analysis, product reviews, and brand reputation management. Sentiment analysis aims to determine the sentiment or emotional tone expressed in a piece of text, which can be positive, negative, or neutral. This information is invaluable for businesses, organizations, and individuals looking to gain insights into public opinion and make informed decisions.

The rapid growth of the internet and the proliferation of social media platforms have led to an explosion of text data containing opinions and sentiments. This wealth of data has created a strong demand for automated methods that can analyze and interpret sentiments at scale. Machine learning techniques have emerged as a powerful tool for sentiment analysis, allowing us to automatically classify and quantify sentiments in large volumes of text.

Furthermore, we will discuss the evaluation metrics commonly used to assess the performance of sentiment analysis models, such as accuracy, precision, recall, F1-score, and the challenges associated with building reliable sentiment lexicons and datasets for training and testing. We will also address ethical considerations in sentiment analysis, including biases in data and models, privacy concerns, and the responsible use of sentiment analysis in decision-making processes.

Finally, we will highlight some recent advancements and emerging trends in sentiment analysis, such as domain adaptation, multi-modal sentiment analysis (combining text with other data sources like images and audio), and the integration of sentiment analysis into real-time applications. We will conclude the survey by summarizing the key takeaways and discussing potential future directions in sentiment analysis research and its continued evolution in the context of machine learning and NLP.

Sentiment analysis is a machine learning tool which is used for analyze the texts for polarity from positive to negative. Machine automatic learn how to analyze the sentiment of the human without the human input or interruption. Nowadays social media is a part of the people’s life; people use social media for give their review over some political field, movie review or marketing area. There are many social media sites like Twitter, Facebook, Instagram etc. They use this social media sites as the medium to express their view on many topics. So, sentiment analysis analyzes the text which inputted by any person from the different country by using the training data set it will analyze the sentiment of that particular text by knowing the emotion of that people. The application of the sentiment analysis very broad and powerful like Expedia Canada; Canadian take the advantage of sentiment analysis when they notice that people are giving negative comments on the music used by their television channel. Rather than chalking by negative comment, Expedia manages to take advantage of that negative comment and air all new soulful music in their channel.

several research papers that studied Twitter’s data classification and analysis for different purposes were surveyed to investigate the methodologies and approaches utilized for text classification. The authors of this research paper aim to obtain open-source datasets then conduct text classification experiments using machine learning approaches by applying different classification algorithms, i.e., classifiers. The authors utilized several classifiers to classify texts of two versions of datasets. The first version is unbalanced datasets, and the second is balanced datasets. The authors then compared the classification accuracy for each used classifier on classifying texts of both datasets.

## 2. Literature Survey

As social media websites have attracted millions of users, these websites store a massive number of texts generated by users of these websites. Researchers were interested in investigating these metadata for search purposes. In this section, a number of research papers that explored the analysis and classification of Twitter metadata were surveyed to investigate different text classification approaches [1] and the text classification results.

Researchers of [2] investigated the user’s gender of Twitter. Authors noticed that many Twitter users use the URL section of the profile to point to their blogs, and the blogs provided valuable demographic information about the users. Using this method, the authors created a corpus of about 184000 Twitter users labeled with their gender. Then authors arranged the dataset for experiments as following: for each user; they specify four fields; the first field contains the text of the tweets and the remaining three fields from the user’s profile on Twitter, i.e., full name, screen name, and description. After that, the authors conducted the experiments and found that using all of the dataset fields while classifying Twitter user’s gender provides the best accuracy of 92%. Using tweets text only for classifying Twitter user’s gender provides an accuracy of 76%. In, the authors used Machine Learning approaches for Sentiment Analysis. Authors constructed a dataset consisting of more than 151000 Arabic tweets labeled as "75,774 positive tweets and 75,774 negative tweets". Several Machine Learning Algorithms were applied, such as Naive Bayes (NB), AdaBoost, Support vector machine (SVM), ME, and Round Robin (RR). The authors found that RR provided the most accurate results on classifying texts, while AdaBoost classifier results were the least accurate results. A study by interested as well in Sentiment Analysis of Arabic texts. The authors constructed the Arabic Sentiment Tweets Dataset ASTD, which consists of 84,000 Arabic tweets. The number of tweets remaining after annotation was around 10,000 tweets. The authors applied machine learning approaches using classifiers on the collected dataset. They reported the following: (1) The best classifier applied on the dataset is SVM, (2) Classifying a balanced set is challenging compared to the unbalanced set. The balanced set has fewer tweets than the unbalanced set, which may negatively affect the classification's reliability. In [3], the author investigated the effects of applying preprocessing methods before the sentiment classification of the text. The authors used classifiers and five datasets to evaluate the preprocessing method's effects on the classification. Experiments were conducted, and researchers reported the following findings: Removing URL has no much effect, removing stop words have a slight effect, Removing Numbers have no effect, Expanding Acronym improved the classification performance, and the same preprocessing methods have the same effects on the classifier's performance, NB and RF classifiers showed more sensitivity than LR and SVM classifiers. In conclusion, the classifier's performance for sentiment analysis was improved after applying preprocessing methods. A study by [30] investigated Twitter geotagged data to construct a national database of people's health behavior. The authors compared indicators generated by machine learning algorithms to indicators generated by a human. The authors collected around 80 million geotagged tweets. Then Spatial Join procedures were applied, and 99.8% of tweets were successfully linked. Then tweets were processed. After that, machine learning approaches were used and successfully applied in classifying tweets into happy and not happy with high accuracy. In [4] explored classifying sentiments in movie reviews. The authors constructed a dataset of 21,000 tweets of movie reviews. Dataset split into train set and test set. Preprocessing methods applied, then two classifiers, i.e., NB and SVM, were used to classify tweets text into positive or negative sentiment. The authors found that better accuracy achieved using SVM of 75% while NB has 65% accuracy. Researchers used Machine Learning methods and Semantic Analysis for analyzing tweet's sentiments. Authors labeled tweets in a dataset that consists of 19340 sentences into positive or negative. They applied preprocessing methods after that features were extracted; authors applied Machine Learning approaches, i.e., Naïve Bayes, Maximum Entropy, and Support Vector Machine (SVM) classifiers after that Semantic Analysis were applied. The authors found that Naïve Bayes provided the best accuracy of 88.2, the next SVM of 85.5, and the last is Maximum entropy of 83.8. The authors reported as well that after applying Semantic Analysis, the accuracy increased to reach 89.9. In [5], the authors analyzed sentiments by utilizing games. Authors introduced TSentiment, which is a web-based game. TSentiment used for emotion identification in Italian tweets. TSentiment is an online game in which the users compete to classify tweets in the dataset consists of 59,446 tweets. Users first must evaluate the tweet's polarity, i.e., positive, negative, and neutral, then users have to select the tweet's sentiment from a pre-defined list of 9 sentiments in which 3 sentiments identified for the positive polarity, 3 sentiments identified for negative polarity. Neutral polarity is used for tweets that have no sentiment expressions. This approach for classifying tweets was effective.

## 3. Problem Statement

In this work, the authors implemented and evaluated different classifiers in classifying the sentiment of the tweets. It's by utilizing RapidMiner software. Classifiers were applied on both balanced and unbalanced datasets. Classifiers used are Decision Tree, Naïve Bayes, Random Forest, K-NN, ID3, and Random Tree.

## 4. Proposed System

In this section, the dataset is described as well as the settings and evaluation techniques are used in the experiments have been discussed. The prediction for the tweet category is tested twice—the first time on an unbalanced data set and the second time on a balanced dataset as below.

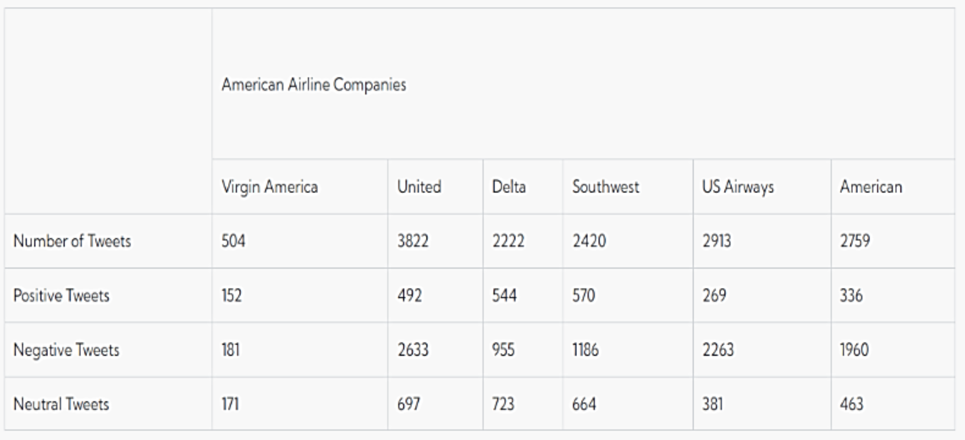
Experiments on the unbalanced dataset: Decision Tree, Naïve Bayes, Random Forest, K-NN, ID3, and Random Tree classifiers were applied on six unbalanced datasets.

Experiments on the balanced dataset: In this experiment, the challenges related to unbalanced datasets were tackled by manual procedures to avoid biased predictions and misleading accuracy. The majority class in each dataset almost equalized with the minority classes, i.e., many positive, negative, and neutral, practically the same in the balanced dataset as represented.

### 4.1 Dataset Description

We obtained a dataset from Kaggle, one of the largest online data science communities in this work. It consists of more than 14000 tweets, labeled either (positive, negative, or neutral). The dataset was also split into six datasets; each dataset includes tweets about one of six American airline companies (United, Delta, Southwest, Virgin America, US Airways, and American). Firstly, we summarized the details about the obtained datasets, as illustrated in Table 1 below.

Table 1: Summary of obtained Dataset



| | American Airline Companies | | | | | |
|---|---|---|---|---|---|---|
| | Virgin America | United | Delta | Southwest | US Airways | American |
| Number of Tweets | 504 | 3822 | 2222 | 2420 | 2913 | 2759 |
| Positive Tweets | 152 | 492 | 544 | 570 | 269 | 336 |
| Negative Tweets | 181 | 2633 | 955 | 1186 | 2263 | 1960 |
| Neutral Tweets | 171 | 697 | 723 | 664 | 381 | 463 |

### 4.2. Dataset Cleansing

In this section, the authors described the followed procedure in the dataset preparation. The authors utilized RapidMinor software for tweet classification. Authors followed the methods described below:

- Splitting the dataset into a training set and test set.
- Loading the dataset, i.e., excel file into RapidMinor software using Read Excel operator.
- Applying preprocessing by utilizing the below operators.
  - Transform Cases operator to transform text to lowercase.
  - Tokenize operator to split the text into a sequence of tokens.

- Filter Stop words operator to remove stop words such as: is, the, at, etc.
- Filter Tokens (by length) operator: to remove token based on the length, in this model, minimum characters are 3, and maximum characters are 20 any other tokens that don't match the rule will be removed.
- Stem operator: to convert words into base form.

### 4.3. Dataset Training

Each of the datasets was divided into two-part. The first part contains 66% of the total number of tweets of the data set, and it is used to train the machine to classify the data under one attribute, which is used to classify the tweets to either (positive or Negative or Neutral). The remaining 34% of tweets were used to classify tweets' attribute to (positive or Negative or Neutral), i.e., test set.

### 4.4. Dataset Classifying

In this section, the authors described the steps in the tweet's classification techniques.

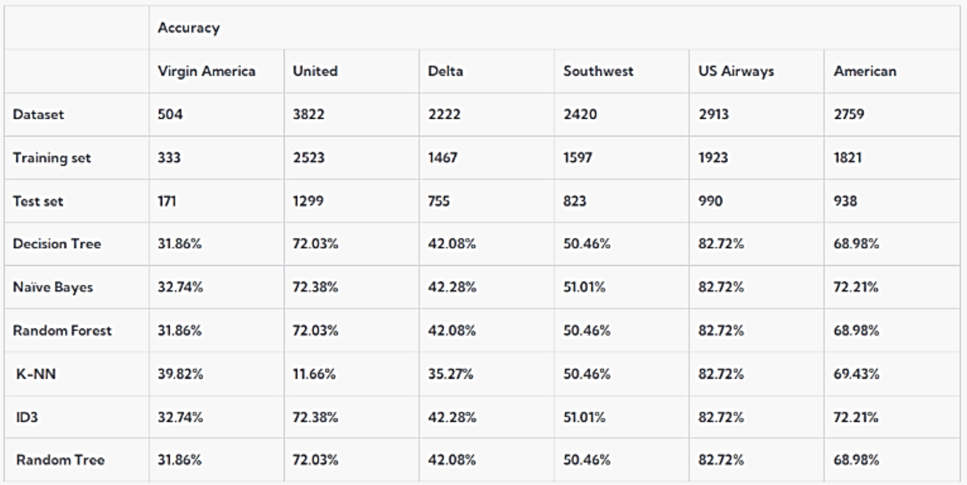
- Set Role operator is used to allow the system to identify sentiment as the target variable,
- Select Attributes operator is used to removing any attribute which has any missing values.
- Then in the validation operator, the dataset is divided into two parts (training and test). We used Two-thirds of the dataset to train the dataset and the last one-third to evaluate the model.
- Different machine learning algorithms are used for training the dataset (Decision Tree, Naïve Bayes, Random Forest, K-NN, ID3, and Random Tree).
- For testing the model, the Performance operator utilized to measure the performance of the model.

## 5. Result

This section presented the experiment results in terms of accuracy level of prediction for each classifier on both types of datasets (balanced, unbalanced) and a comparison between the two experiments.

Experiment results for an unbalanced dataset
Figure 2 and Table 2 present the accuracy results of the utilized classifiers on the datasets.

Table 2: Accuracy results on unbalanced dataset



| | Accuracy | | | | | |
|---|---|---|---|---|---|---|
| | Virgin America | United | Delta | Southwest | US Airways | American |
| Dataset | 504 | 3822 | 2222 | 2420 | 2913 | 2759 |
| Training set | 333 | 2523 | 1467 | 1597 | 1923 | 1821 |
| Test set | 171 | 1299 | 755 | 823 | 990 | 938 |
| Decision Tree | 31.86% | 72.03% | 42.08% | 50.46% | 82.72% | 68.98% |
| Naïve Bayes | 32.74% | 72.38% | 42.28% | 51.01% | 82.72% | 72.21% |
| Random Forest | 31.86% | 72.03% | 42.08% | 50.46% | 82.72% | 68.98% |
| K-NN | 39.82% | 11.66% | 35.27% | 50.46% | 82.72% | 69.43% |
| ID3 | 32.74% | 72.38% | 42.28% | 51.01% | 82.72% | 72.21% |
| Random Tree | 31.86% | 72.03% | 42.08% | 50.46% | 82.72% | 68.98% |

In some datasets, the classifier's accuracy results were very high, while it was low in others. All classifier's performance on the US airways dataset and United dataset provided the best accuracy due to the dataset's size, which was the largest. Naïve Bayes classifier, Decision Tree, and ID3 were mostly better than other classifiers and were given almost the same accuracy level. The classifiers with Virgin America dataset reported the lowest accuracy level due to the dataset's size, which is very small.

## 6. Conclusion

Social media websites are gaining very big popularity among people of different ages. Platforms such as Twitter, Facebook, Instagram, and Snapchat allowed people to express their ideas, opinions, comments, and thoughts. Therefore, a huge amount of data is generated daily, and the written text is one of the most common forms of the generated data. Business owners, decision-makers, and researchers are increasingly attracted by the valuable and massive amounts of data generated and stored on social media websites. Sentiment Analysis is a Natural Language Processing field that increasingly attracted researchers, government authorities, business owners, services providers, and companies to improve products, services, and research. In this research paper, the authors aimed to survey sentiment analysis approaches. Therefore, 16 research papers that studied Twitter's text classification and analysis were surveyed. The authors also aimed to evaluate different machine learning algorithms used to classify sentiment to either positive or negative, or neutral. This experiment aims to compare the efficiency and performance of different classifiers that have been used in the sixteen papers that are surveyed. These classifiers are (Decision Tree, Naïve Bayes, Random Forest, K-NN, ID3, and Random Tree). Besides, the authors investigated the balanced dataset factor by applying the same classifiers twice on the dataset, one on the unbalanced and the other, after balancing the dataset. The targeted dataset included six datasets about six American airline companies (United, Delta, Southwest, Virgin America, US Airways, and American); it consists of about 14000 tweets. The authors reported that the classifier's accuracy results were very high in some datasets while low in others. The authors indicated that the dataset size was the reason for that. On the balanced dataset, the Naïve Bayes classifier, Decision Tree, and ID3 were mostly better than other classifiers and have given the almost same level of accuracy. The classifiers with Virgin America dataset reported the lowest level of accuracy due to its small size. On the unbalanced dataset, results show that the Naive Byes and ID3 gave a better level of accuracy than other classifiers when it's applied on the balanced datasets. While (K-NN, Decision Tree, Random Forest, and Random Tree) gave a better understanding of the unbalanced datasets.

## References

1. S.A. Salloum, C. Mhamdi, B. Al Kurdi, K. Shaalan, "Factors affecting the Adoption and Meaningful Use of Social Media: A Structural Equation Modeling Approach," International Journal of Information Technology and Language Studies, 2(3), 2018.
2. M. Alghizzawi, S.A. Salloum, M. Habes, "The role of social media in tourism marketing in Jordan," International Journal of Information Technology and Language Studies, 2(3), 2018.
3. S.A. Salloum, W. Maqableh, C. Mhamdi, B. Al Kurdi, K. Shaalan, "Studying the Social Media Adoption by university students in the United Arab Emirates," International Journal of Information Technology and Language Studies, 2(3), 2018.
4. S.A. Salloum, M. Al-Emran, S. Abdallah, K. Shaalan, Analyzing the arab gulf newspapers using text mining techniques, 2018, doi:10.1007/978-3-319-64861-3_37.
5. F.A. Almazrouei, M. Alshurideh, B. Al Kurdi, S.A. Salloum, Social Media Impact on Business: A Systematic Review, 2021, doi:10.1007/978-3-030-58669-0_62.
6. Alshurideh et al., "Understanding the Quality Determinants that Influence the Intention to Use the Mobile Learning Platforms: A Practical Study," International Journal of Interactive Mobile Technologies (IJIM), 13(11), 157–183, 2019.
7. S.A. Salloum, K. Shaalan, Adoption of E-Book for University Students, 2019, doi:10.1007/978-3-319-99010-1_44.
8. S.A. Salloum, M. Alshurideh, A. Elnagar, K. Shaalan, "Mining in Educational Data: Review and Future Directions," in Joint European-US Workshop on Applications of Invariance in Computer Vision, Springer: 92–102, 2020.
9. K.S.A. Wahdan, S. Hantoobi, S.A. Salloum, K. Shaalan, "A systematic review of text classification research based ondeep learning models in Arabic language," Int. J. Electr. Comput. Eng, 10(6), 6629–6643, 2020.