



# IJRASET

International Journal For Research in  
Applied Science and Engineering Technology



---

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

---

**Volume: 9      Issue: VI      Month of publication: June 2021**

**DOI: <https://doi.org/10.22214/ijraset.2021.35807>**

**[www.ijraset.com](http://www.ijraset.com)**

**Call:  08813907089**

**E-mail ID: [ijraset@gmail.com](mailto:ijraset@gmail.com)**

# Sentiment Analysis on Twitter Airline Data

Kirti Jain<sup>1</sup>, Jaishree<sup>2</sup>, Kajal Sharma<sup>3</sup>, Muskan Jain<sup>4</sup>, Ishika Singh<sup>5</sup>

<sup>1, 2, 3, 4, 5</sup>Department of Computer Science & Engineering, Inderprastha Engineering College, Ghaziabad, Uttar Pradesh, India

**Abstract:** Sentiment analysis, also known as sentiment mining, is a submachine learning task where we want to determine the overall sentiment of a particular document. With machine learning and natural language processing (NLP), we can extract the information of a text and try to classify it as positive, neutral, or negative according to its polarity. In this project, We are trying to classify Twitter tweets into positive, negative, and neutral sentiments by building a model based on probabilities. Twitter is a blogging website where people can quickly and spontaneously share their feelings by sending tweets limited to 140 characters. Because of its use of Twitter, it is a perfect source of data to get the latest general opinion on anything.

**Keywords:** Sentiment Analysis, Machine Learning, Airline dataset; Naïve-Bayes classifier

## I. INTRODUCTION

We live in a society where the amount of Internet text data is increasing rapidly. Numerous companies are trying to use this data avalanche to understand people's views on their products. Online social media platforms have large repositories of user-generated content, which can provide unique opportunities to stimulate the emotional pulse of national and global communities. Social media contains a large amount of unstructured text information, and manual analysis of such a large amount of data is impractical. Numerous social networking sites allow users to contribute, change, and rate content. Some examples include blogs, forums, product review sites, and social media, such as Twitter (<http://twitter.com/>). Twitter is a blog site that provides the ability to analyze and express emotions. Previous research has shown that geographic and daily data can produce continuous, weekly, and seasonal positive and negative impact patterns.

## II. RELATED WORK

The researchers [2] presented a "Web-Based Opinion Mining system" for hotel evaluations. The research presented an assessment methodology for online user evaluations and comments to assist quality controls in hotel management systems. It can recognize and retrieve online reviews and works with German reviews. It features a multi-topic domain and is based on multi-polarity classification; the system may identify neutral expressions such as "don't know" to "classify emotion polarity that as neutral." It is the most obvious flaw in not dealing with some scenarios in multi-topic segments.

examined product reviews for mobile gadgets in [3]. Machine learning was used, and the ML system [1] tested the classification accuracy of the Naive Bayes technique. Furthermore, the research provided a favourable assessment of the product quality and market situation. This research used three ML algorithms (Naive Base Classifier, K-nearest neighbour, and random forest) to calculate the accuracy of the sentiment. The random forest algorithm improves the performance of the classifier. The paper [4] analyses news sentiments and blogs. It splits prior work in the context of their specific task into two categories. The first category is concerned with approaches for automatically building sentiment lexicons, while the second is concerned with systems that assess sentiment across full documents.

Corpus-based techniques examine the inclusion of words based on big groups of text [5] or look for context dependent labels while taking into account local limits [6]. People have also looked at studying knowledge encoded in WordNet as relations and glosses. The majority of subjectivity detection research for sentiment polarity classification assumes that the input documents are opinionated. Many tools and applications need a choice on whether a particular document contains subjective information or not. Previous research in this area [7] investigated the effects of adjective orientation and gradeability on sentence subjectivity. The target told us whether or not a certain statement is subjective based on the adjectives in that sentence.

The research on YouTube opinion mining undertaken [8] for examining how social media may be used to radicalize a person can identify information and interaction aimed at radicalizing persons with no or little apparent prior interest in violent Jihadism. Their work investigates a truly fruitful method. They gathered a significant dataset from a YouTube collection that was identified as potentially having an advanced objective. It also investigates the subjects mentioned and the polarity of sentiment toward these issues.

Using labelled data obtained from online sources, the researchers used supervised learning algorithms to tackle the related challenge of identifying sentiment polarity in reviews. Interestingly, our research on this challenge suggests that people may not always have the best intuition when it comes to selecting discriminating terms. While they experiment with a variety of features in prior research [9], their primary focus was not feature engineering.

The trend of sentiment analysis research is not limited to no. It enables the suggested approaches, applications, or algorithms to be more compatible with comprehending meaning and features to increase accuracy and performance.

In [10], the focus is on airline-related tweets since customers share their experiences on Twitter through their tweets, and our analyzer helps the airline business improve their services by keeping a watch on people's feelings and overcoming their defects.

### III.DATASET

In our proposed study, we used a dataset of airline tweets from the Kaggle data source. Here is the link for a more detailed explanation: <https://www.kaggle.com/crowdflower/twitter-airline-sentiment>

### IV.METHODOLOGY

Proposed work includes some modules where flow of the work is explained through proposed architecture. Step involved in proposed methodology are as: Functionality and Design:

- 1) *Data Acquisition* - It is the process of sampling signals which measure real world physical conditions and convert the resulting samples into digital numeric values that can be manipulated by a computer.
- 2) *Human Labelling* - It enables you to label images in your training dataset at scale. Well-labeled content results in better training data that results in more accurate predictions from your model.
- 3) *Feature Extraction*- It is a part of the dimensionality reduction in which an initial set of the raw data is divided and then reduced to more manageable groups.

Classification –

#### A. Step 1: Data Collection and Preparation

Data can be collected either from the user or from the existing system. In proposed code we have used Airline Tweet Dataset from Kaggle data repository.

#### B. Step 2: Data Pre-processing

- 1) *Data Cleaning*: Data cleaning is done to remove redundant data, irrelevant tweets & unwanted link.
- 2) *Lemmatization*: Stanford Lemmatize Library is used. Lemmatization refers to vocabulary and morphological words. It removes inflected word and concentrate on words involved in dictionary.
- 3) *Tokenization*: Here, sensitive data is replaced with non-sensitive equivalent data and are represented as tokens.

#### C. Step 3: Proposed Estimation

Ambiguity based polarity estimation: To check polarity of ambiguity word whether the word is positive, negative or neutral.

This work will initially classify Tweets in three categories which are positive, negative and ambiguous. Example of ambiguous Tweets; Ambiguous tweet: “This is good airline service but comes with high tariff. They do not provide good service. Best part of airline is they are always available. “Positive tweet: “This one is the best airline comes with awesome services and facility. 100 % recommendation for new once” Negative tweet: “Worst services experience. Air hostess don’t know about their responsibilities”

This work will investigate ambiguous tweets and observe sentiments of every word. After observing sentiments for every person, it also checks sentiment of previous and next word. In case of ambiguous word it fixes the word polarity based on occurrence and whether they are positive or negative.

The complete Tweets will be forwarded to Naïve Bayes classifier.

#### D. Step 4: Classifier

Naïve Bayes classifier: It is used for large volume of dataset. Naïve Bayes classifier supposes that features are independent from values of given class labels. It will apply unigram and n-gram and concludes results based on both techniques.

E. Step 5: Tweets Classification

1) Sentiment Calculation: polarity of sentiments is calculated.

V. RESULT

The following pie charts and graphs show the result of our project: -

First, we have plotted a pie chart for the number of tweets for each airline.

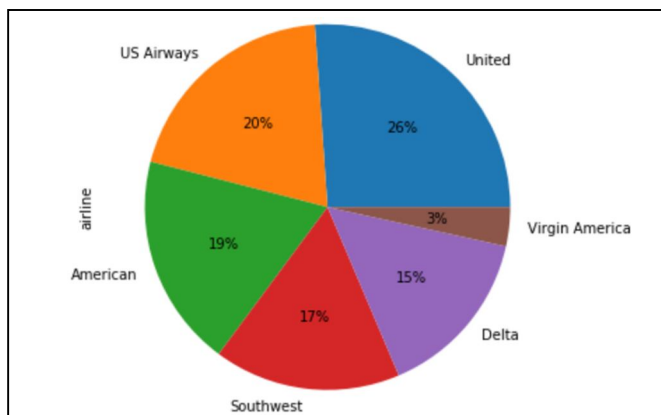


Fig. 1 pie chart for number of tweets for each airline.

The following pie chart shows the distribution of sentiments across all the tweets.

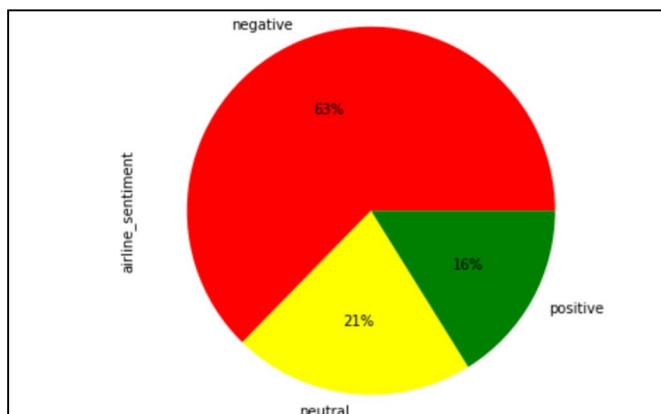


Fig. 2 pie chart for the distribution of sentiments across all the tweets.

The following graph shows the distribution of the sentiments across all the tweets for different airlines.

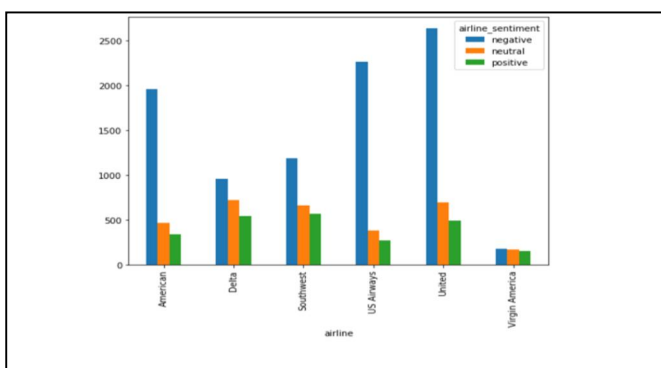


Fig. 3 Graph for the distribution of sentiments across all the tweets for different airlines.



The following graph shows confidence level for negative tweets is higher compared to positive and neutral tweets.

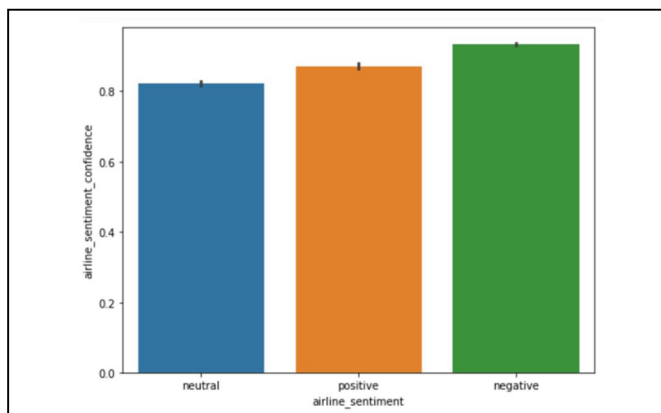


Fig. 4. Graph for the confidence level for negative tweets is higher compared to positive and neutral tweets.

## VI. CONCLUSION

The entire study is built on the Twitter Dataset for Airline Sentiments, where the airline dataset is utilized as the original input, with the data being analysed and pre-processed to determine the sentiments of the words. Based on the sentimental analysis technique, this study provided a sentiment analysis model to notice positive and negative viewpoints of different users. The whole work pre-processes and cleans the data using the Naive Bayes classifier, and the sentiments of airline tweets are calculated on this premise. To avoid the unigram issue of previous work, which takes a single word tweet, Naive Bayes is employed in suggested work, which takes an entire tweet to determine the number of words in the tweet and their polarity.

## VII. FUTURE WORK

For the proposed solution, the following future work is anticipated.

- A. For large datasets, the proposed technique can be implemented utilizing big data technologies.
- B. The proposed technique may be used to categorize tweets in all categories except airlines.
- C. To improve performance, the proposed method may be used with other classification algorithms such as SVM and KNN.

## REFERENCES

- [1] Nisha, J., & Dr.E. Kirubakaran, "M-Learning sentiment analysis with Data Mining Techniques", International Journal of Computer Science and Telecommunications, Volume 3, Issue 8, 2012.
- [2] Kasper, W. & Vela, M., "Sentiment analysis for hotel reviews", proceedings of the computational linguistics-applications, Jacharanka Conference, 2011.
- [3] Zhang, L., Hua, K., Wang, H., and Qian, G., "Sentiments reviews for mobile devices products", The 11th International Conference on Mobile Systems and Pervasive Computing (MobiSPC-2014), procedia computer science, Volume 34, 2014.
- [4] Godbole, N., Srinivasaiah, M., and Skiena, S., "Large-Scale Sentiment Analysis for News and Blogs", ICWSM'2007 Boulder, Colorado, USA, 2007.
- [5] Turney, P., "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews". In Proc. of the ACL, 2002.
- [6] Argamon-Engelson, S., Koppel, M., and Avneri, G., "Style-based text categorization: What newspaper am I reading?", In Proc. of the AAAI Workshop on Text Categorization, pages 1-4, 1998.
- [7] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S., Weld, D. and Yates, A., "Unsupervised named-entity extraction from the web: An experimental study", Artificial Intelligence, 165(1):91-134, 2005.
- [8] Jin, W., & HOH. H., "A novel lexicalized HMM-based learning framework for web opinion mining", Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Quebec, Canada, ACM: 465-472, 2009.
- [9] Brody, S., & Elhadad, N., "An unsupervised aspect-sentiment model for online reviews", Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California, Association for Computational Linguistics: 804-812, 2010.
- [10] Jain, K., Singh, A., & Yadav, A. (2019). Sentiment analysis using twitter data. International Research Journal of Engineering and Technology (IRJET), 6(5), 4746-4751.



10.22214/IJRASET



45.98



IMPACT FACTOR:  
7.129



IMPACT FACTOR:  
7.429



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24\*7 Support on Whatsapp)