SENTIMENT ANALYSIS ON TWEETS OF TWITTER(X)

UDAY KUMAR CHINNI[1] K. SOMA SEKHAR[2] S. BHANU PRASAD[3]

STUDENTS OF
BACHELOR'S OF TECHNOLOGY DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
AT
DADI INSTITUTE OF ENGINEERING AND TECHNOLOGY (AUTONOMOUS)
udaykumarchinni2706@gmail.com

Abstract

The increase in the usage of internet is greatly increased, and the rapid advancements in web technology has led to an exponential surge in the volume of data available on the internet. A substantial amount of unstructured data is being generated in real-time. The internet is a dynamic platform for online learning, opinion sharing, exchanging ideas and among the various platforms Twitter (X), Facebook have witnessed increased popularity within the last decade. Individuals use these platforms to share their opinion or perspective, engage in discussions with various communities in the world and can be used to deliver a message to the world.

To understand the opinion of each person on a particular discussion a considerable amount of research has been dedicated to the field of sentiment analysis, particularly concerning Twitter (X) data. Twitter (X) has emerged as a prominent platform in the digital landscape, playing a pivotal role in facilitating communication, information dissemination, and engagement among users worldwide. Sentiment analysis is of great importance due to the unstructured data and highly differing nature of opinions expressed in tweets, which are generally classified as either positive, negative, or occasionally neutral on the particular topic mentioned in a tweet. This project is based on the domain of sentiment analysis applied to twitter data and to provide insights into how sentiment analysis can be employed to tweets of every user on a particular topic, which helps one to understand the large discussion held with simple steps, where the information i.e., the tweet is of complex and of varying from one to another. Using various machine learning algorithms like Logistic Regression, LightGBM, XGBoost, Support Vector Machine we will implement the project of sentiment analysis on Twitter (X) data and based on the accuracy one can use the most accurate algorithm.

Keywords:
Sentiment Analysis, Twitter Data, Web Scraping, Machine Learning, Unstructured Data, Positive Comment, Negative Comment, Neutral Comment.

1. Introduction

In the contemporary era of the Internet, the manner in which individuals articulate their viewpoints and sentiments has undergone a transformation, primarily occurring through avenues such as blog posts, online forums, product review platforms, and social media networks. Presently, millions of individuals engage in platforms like Facebook, Twitter, and Google Plus to convey their emotions, share opinions, and discuss their daily experiences. Within online communities, there exists an interactive medium where consumers both inform and influence others through discussions and forums. The proliferation of social media has resulted in a substantial volume of sentiment-laden data, comprising tweets, status updates, blog entries, comments, and reviews. Furthermore, social media platforms offer businesses an opportunity to engage with their customer base for advertising purposes. A significant portion of the populace heavily relies on user-generated content online for decision-making processes, such as researching product reviews and discussing potential purchases on social media platforms before making informed choices. Given the vast amount of user-generated content available, there is a pressing need for automation, leading to the widespread use of various sentiment analysis techniques. Sentiment analysis serves to inform users about the satisfaction levels associated with products or services before making purchasing decisions. Marketers and businesses utilize sentiment analysis data to tailor their offerings according to user preferences and requirements. Textual information retrieval methods primarily focus on processing, searching, or analyzing factual data, which inherently possesses an objective component. However, there exists another category of textual content expressing subjective characteristics, including opinions, sentiments, appraisals, attitudes, and emotions, constituting the essence of sentiment analysis. The abundance of information available on online platforms like blogs and social networks presents numerous challenging opportunities for the development of new applications, such as predicting item recommendations based on the sentiment analysis of user opinions.

2. Literature Survey

The research community is actively evaluating the significant impact of Twitter applications on various companies today, with a particular focus on the consistent analysis of Twitter sentiment. One key challenge in this analysis lies in the intricate structure of the retrieved data and the diverse nature of speech.

In a comprehensive study, data from two distinct datasets with different characteristics underwent analysis using four classification algorithms and ensemble techniques to enhance reliability. Surprisingly, the tests revealed that the use of a single algorithm slightly outperformed ensemble techniques. Additionally, the analysis concluded that employing 50% of the data as training data yielded results comparable to using 70% of the data for training.

Another aspect of the investigation centred on the analysis of Twitter data related to demonetization. Utilizing the R programming language, graphical plots with word clouds based on tweet analysis were presented. The plotted results led to the conclusion that a considerable majority of individuals expressed acceptance of demonetization compared to those who rejected it.

In the realm of Twitter data studies, a straightforward approach was taken by the researcher, extracting tweets in JSON format and determining tweet polarity using the Python Lexicon Dictionary. On the contrary, a more sophisticated strategy was adopted, leveraging learning approaches to enhance accuracy. Focusing on crypto currency data, Support Vector Machine (SVM) and Naïve Bayes algorithms were applied, revealing that the Naïve Bayes classifier outperformed SVM in accuracy.

In a distinct investigation, the unigram model served as a baseline, compared with experimental models based on features and kernel trees. The results highlighted the superior performance of the kernel tree-oriented model over both unigram and feature-oriented models, while the feature-oriented model exhibited a slight edge over the unigram model.

An unconventional approach was taken, combining a corpus-based approach with a lexicon-based one, a rarity in the predominantly machine learning-focused research landscape.

The researcher delved into public opinion on geographical flood data collected from Twitter, employing the Naïve Bayes algorithm to achieve a 67% accuracy rate. The importance of gathering diverse measures from the public to enhance situational management was emphasized.

A focused analysis aimed to discern the utility of sentiment analysis for both customers and online businesses, exploring the demand and impact of this analytical approach.

The researcher also aimed to discern the emotional responses of viewers to a random television program, collecting remarks from a selection of diverse TV broadcasts. These comments served as data for training and testing a Naive Bayes classifier model, revealing a prevalence of negative tweets over positive ones in terms of polarity.

In 2010, a study delved into the potential of Twitter data in predicting elections. Analysing political sentiment expressed within Twitter's 140-character limit, the research explored the correlation between Twitter activity and election outcomes.

The paper concentrated on sentiment analysis, providing a comparative analysis of various sentiment analysis approaches, offering insights into their strengths, weaknesses, and application domains.

The researcher harnessed social media and news data for sentiment analysis, utilizing Naïve Bayes and Levenshtein methods to categorize sentiments from diverse sources. This approach not only enhanced lead-to-term accuracy but also demonstrated robust performance in real-time news content on social media. Notably, the Levenshtein formula emerged as a swift and efficient means of processing a substantial amount of information with high accuracy levels.

2.1 Feature Extraction Techniques:

The preprocessed dataset possesses numerous distinct characteristics. During the feature extraction process, we identify aspects within the processed dataset. These aspects are subsequently utilized to assess the positive and negative sentiments in a sentence, aiding in gauging individuals' opinions using models such as unigram and bigram.

Machine learning techniques rely on representing the salient features of text or documents for analysis. These essential features are transformed into feature vectors, pivotal for classification tasks. Some examples of features documented in literature include:

2.1.1 Word Presence and Frequencies:

Utilizing unigrams, bigrams, and n-grams along with their frequency counts as features has been explored. However, recent studies, such as Pan et al., have demonstrated improved results by focusing on word presence rather than frequencies.

2.1.2 Parts of Speech Tags:

Identifying parts of speech, such as adjectives, adverbs, and specific groups of verbs and nouns, serves as valuable indicators of subjectivity and sentiment. Syntactic dependency patterns can be generated through parsing or dependency trees.

2.1.3 Opinion Expressions:

In addition to individual words, phrases and idioms conveying sentiments can also be leveraged as features. For example, "cost someone an arm and a leg."

2.1.4 Term Position:

The positioning of a term within a text can significantly influence its impact on the overall sentiment of the text.

2.1.5 Negation Handling:

Negation, though challenging to interpret, plays a crucial role in sentiment analysis. The presence of negation often reverses the polarity of the opinion, as in the example "I am not happy."

2.1.6 Syntactic Patterns:

Patterns such as collocations in syntax are frequently utilized as features to discern subjectivity patterns by numerous researchers. Emotion Recognition Algorithms:

2.2 Per-processing of the Datasets:

In the pre-processing phase of the datasets, tweets serve as repositories of various opinions expressed in diverse manners by different users. The Twitter dataset utilized in this survey has been pre-labeled into two categories: negative and positive polarity. This labeling facilitates the sentiment analysis process, enabling the observation of the impact of different features. The raw data, with its inherent polarity, is prone to inconsistencies and redundancies. Preprocessing of tweets involves the following steps:

- Eliminating all URLs (e.g., www.xyz.com), hashtags (e.g., #topic), and user mentions (@username).
- Addressing spelling errors and managing sequences of repeated characters.
- Substituting emoticons with their corresponding sentiments.
- Removing all punctuation marks, symbols, and numerical digits.
- Filtering out stop words.
- Expanding acronyms using an acronym dictionary.
- Excluding non-English tweets.

2.3 Datasets for Emotion Recognition:

Table.1. Publicly Available Datasets for Twitter

HASH	Tweets	http://demeter.inf.ed.ac.uk	31,861 Pos tweets 64,850 Neg tweets, 125,859 Neu tweets
EMOT	Tweets and Emoticons	http://twittersentiment.appspot.com	230,811 Pos& 150,570 Neg tweets
ISIEVE	Tweets	www.i-sieve.com	1,520 Pos tweets,200 Neg tweets, 2,295 Neu tweets

Columbia univ.dataset	Tweets	Email: apoorv@cs.columbia.edu	11,875 tweets
Sample	Tweets	http://goo.gl/UQvdx	667 tweets
Stanford dataset	Movie Reviews	http://ai.stanford.edu/~amaas/data/sentiment/	50000 movie reviews
Stanford	Tweets	http://cs.stanford.edu/people/alecmgo/ trainingandtestdata.zip	4 million tweets categorized as positive and negative
Spam dataset	Spam Reviews	http://myleott.com/op_spam	400 deceptive and 400 truthful reviews in positive and negative category.
Soe dataset	Sarcasm and nasty reviews	http://nlds.soe.ucsc.edu/iac	1,000 discussions, ~390,000 posts, and some ~ 73,000,000 words

3. Existing System

The prevailing methods for sentiment analysis heavily lean on human evaluations, manual coding, and subjective judgments. However, these conventional approaches frequently yield inconsistent and occasionally erroneous outcomes, constraining our capacity to extract meaningful insights from vast amounts of textual data. Presently, the existing systems can only assess one comment at a time. Moreover, they commonly encounter challenges in precisely evaluating sentiment within a single comment, leading to diminished model accuracy rates.

This reliance on human assessments and manual processes introduces several limitations. Firstly, human evaluations are inherently subjective, influenced by individual biases and interpretations. Consequently, the consistency and reliability of sentiment analysis outcomes are compromised. Additionally, manual coding is labor-intensive and time-consuming, hindering the scalability of sentiment analysis efforts, particularly when dealing with large datasets.

Furthermore, the current systems' struggle to accurately evaluate sentiment in individual comments underscores the need for more sophisticated and robust methodologies. Often, sentiment analysis models may overlook nuanced expressions or context-dependent sentiments, resulting in suboptimal performance. Addressing these limitations requires advancements in natural language processing (NLP) techniques, including the development of more nuanced sentiment analysis algorithms capable of comprehensively understanding and contextualizing textual data.

By overcoming the constraints of traditional methods and enhancing the accuracy and efficiency of sentiment analysis, researchers and practitioners can unlock the full potential of textual data for valuable insights across various domains, including marketing, customer feedback analysis, and social media monitoring.

4. Proposed System

Our proposed system represents a paradigm shift in sentiment analysis, harnessing the power of machine learning algorithms to provide a data-driven and impartial solution. Our primary aim is to deliver precise categorization of comments into Positive, Negative, and Neutral sentiments, while also identifying Unidentified comments. By leveraging advanced algorithms, we aim to overcome the limitations of traditional methods, ensuring greater accuracy and reliability in sentiment analysis.

This innovative system holds immense promise in empowering businesses and individuals alike with actionable insights derived from sentiment trends within comments. By offering a nuanced understanding of sentiment dynamics, our system enables informed decision-making, facilitating strategic actions based on sentiment analysis outcomes.

One of the key strengths of our system lies in its ability to visualize data intuitively through graphs. By presenting sentiment distributions in a clear and comprehensive manner, our system enhances comprehension and insight generation. These visual representations not only aid in understanding the sentiment landscape but also streamline the process of extracting actionable insights from comment data.

Through our system, users can gain valuable insights into the sentiments expressed within social networking applications. Whether analyzing customer feedback, monitoring brand perception, or gauging public opinion, our system provides a powerful tool for understanding sentiment dynamics and driving informed decision-making.

Overall, our groundbreaking approach to sentiment analysis offers a transformative solution that empowers users to extract meaningful insights from comment data, enabling proactive and strategic decision-making in various domains.

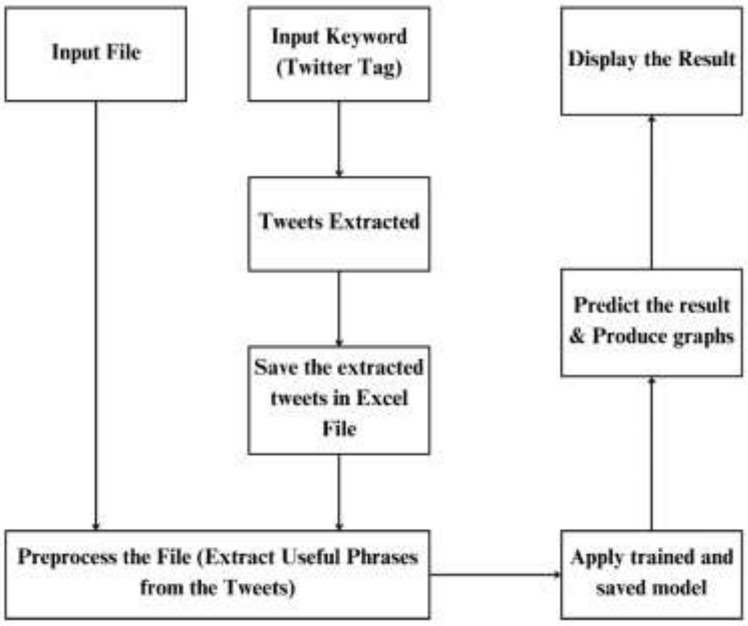

Input File
Input Keyword (Twitter Tag)
Display the Result
Tweets Extracted
Predict the result & Produce graphs
Save the extracted tweets in Excel File
Preprocess the File (Extract Useful Phrases from the Tweets)
Apply trained and saved model

Fig.1. Sentiment Analysis Architecture

6. Machine Learning Approaches

Machine learning plays a pivotal role in sentiment analysis, primarily employing classification techniques to categorize text into different classes. Broadly, there are two types of machine learning techniques:

6.1 Unsupervised Learning:
This method operates without predefined categories and does not provide correct targets. Instead, it relies on clustering algorithms to identify patterns and group similar data points together.

6.2 Supervised Learning:
In contrast, supervised learning relies on labeled datasets, where the model is provided with correct labels during training. These labeled datasets enable the model to learn meaningful patterns and make informed decisions during classification tasks.

The efficacy of both these learning methods hinges on the selection and extraction of relevant features crucial for sentiment detection.

Supervised classification is the dominant approach in sentiment analysis using machine learning techniques. It involves two main sets of data: the training set and the test set. Various machine learning algorithms have been developed to classify tweets into different sentiment classes. Notable techniques include Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM), all of which have demonstrated considerable success in sentiment analysis tasks.

The machine learning approach applicable to sentiment analysis mainly belongs to supervised classification. In a machine learning technique, two sets of data are needed:
1. Training Set
2. Test Set.

The machine learning process typically begins with the collection of a training dataset. Next, a classifier is trained on this dataset to learn patterns and relationships between features and sentiment labels. The selection of features is a critical decision in this process, as they determine how documents are represented and influence the performance of the classification model.

Commonly used features in sentiment classification include term presence and frequency, part-of-speech information, negations, and opinion words and phrases.

Supervised techniques such as SVM, Naive Bayes, and Maximum Entropy are frequently employed due to their effectiveness in sentiment analysis tasks. However, in scenarios where obtaining a labeled dataset is challenging, semi-supervised and unsupervised techniques are proposed to classify unlabeled data based on inherent patterns and structures within the dataset.

7. Machine Learning on Sentiment Analysis

7.1 Naive Bayes:
Naive Bayes operates as a probabilistic classifier, learning patterns from categorized documents. It evaluates documents by comparing their contents with a list of words to assign them to the appropriate class. The classification process involves calculating probabilities based on the occurrence of features within the document. Parameters such as P(c) and P(f|c) are estimated using maximum likelihood techniques, with smoothing applied to unseen features. Implementation of Naive Bayes for training and classification tasks can be facilitated using the Python NLTK library.

$$|C^* = \arg mac_c P_{NB}(c \mid d)$$

$$P_{NB}(c \mid d) = \frac{(P(c))\sum_{i=1}^{m} p(f \mid c)^{n_{i(d)}}}{P(d)}$$

From the above equation, “f” is a “feature”, count of feature (fi) is denoted with $n_i(d)$ and is present in d which represents a tweet. Here, m denotes no. of features.

7.2 Maximum Entropy:
The Maximum Entropy Classifier does not make assumptions about feature relationships within the dataset. Instead, it aims to maximize system entropy by estimating the conditional distribution of class labels. Unlike Naive Bayes, Maximum Entropy handles overlapping features and resembles logistic regression in its approach to finding class distributions. The model representation includes the calculation of conditional probabilities based on feature weights represented by the weight vector λ_i.

$$P_{ME}(c \mid d, \lambda) = \frac{\exp[\sum_i \lambda_i f_i(c,d)]}{\sum_c \exp[\sum_i \lambda_i f_i(c,d)]}$$

Where c is the class, d is the tweet and λ_i is the weight vector. The weight vectors decide the importance of a feature in classification.

7.3 Support Vector Machine:
Support Vector Machine (SVM) analyzes data, delineates decision boundaries, and utilizes kernels for computation within the input space. It operates on two sets of vectors, classifying each data vector into a class. SVM aims to maximize the margin between classes, defining the classifier's decision boundary. This margin optimization reduces ambiguous classifications. SVM supports both classification and regression tasks, contributing to statistical learning theory and aiding in the precise recognition of relevant factors essential for successful understanding.

8. Sentiment Analysis Prediction

Sentiment analysis on tweets from Twitter involves several key steps to extract and analyze sentiment from the textual data. Initially, data is collected either through Twitter's API or other sources, often based on specified keywords, hashtags, or user handles. Preprocessing of the collected tweets is then carried out, involving tasks like removing URLs, mentions, and special characters, as well as tokenization and normalization of text. Feature extraction follows, where relevant features such as word frequency, parts of speech tags, and sentiment lexicons are selected. Next, a machine learning model is trained using labeled data, typically employing algorithms like Naive Bayes or Support Vector Machines. The trained model is evaluated to assess its performance in predicting sentiment accurately. Once validated, the model is deployed to analyze sentiment in new tweets, assigning labels such as positive, negative, or neutral. Post-processing techniques may be applied for further refinement, and results can be visualized using charts or graphs to facilitate interpretation. This systematic approach enables analysts to gain insights into public opinion, brand perception, and other valuable insights from Twitter data.

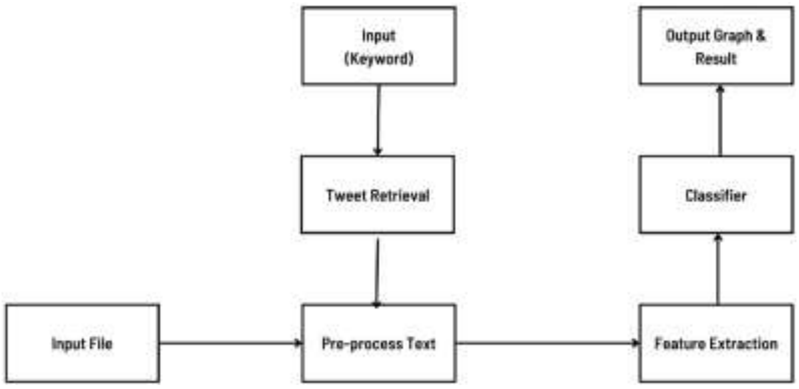
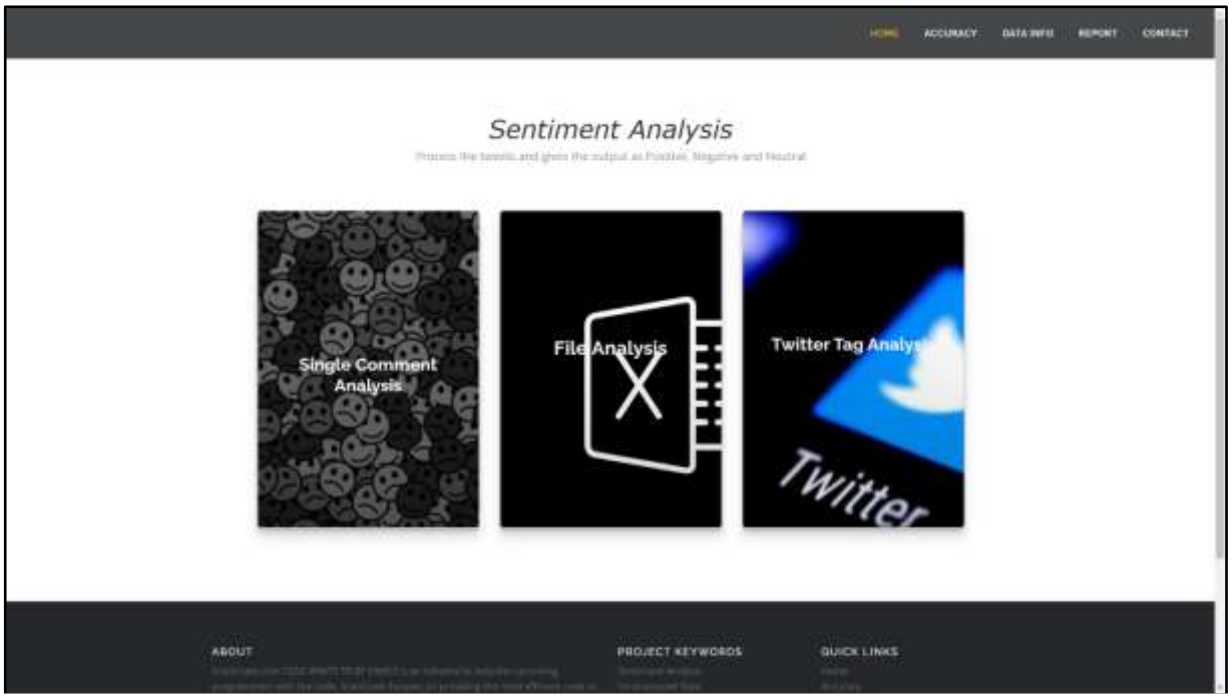


Fig.2. Block Diagram for the Proposed System

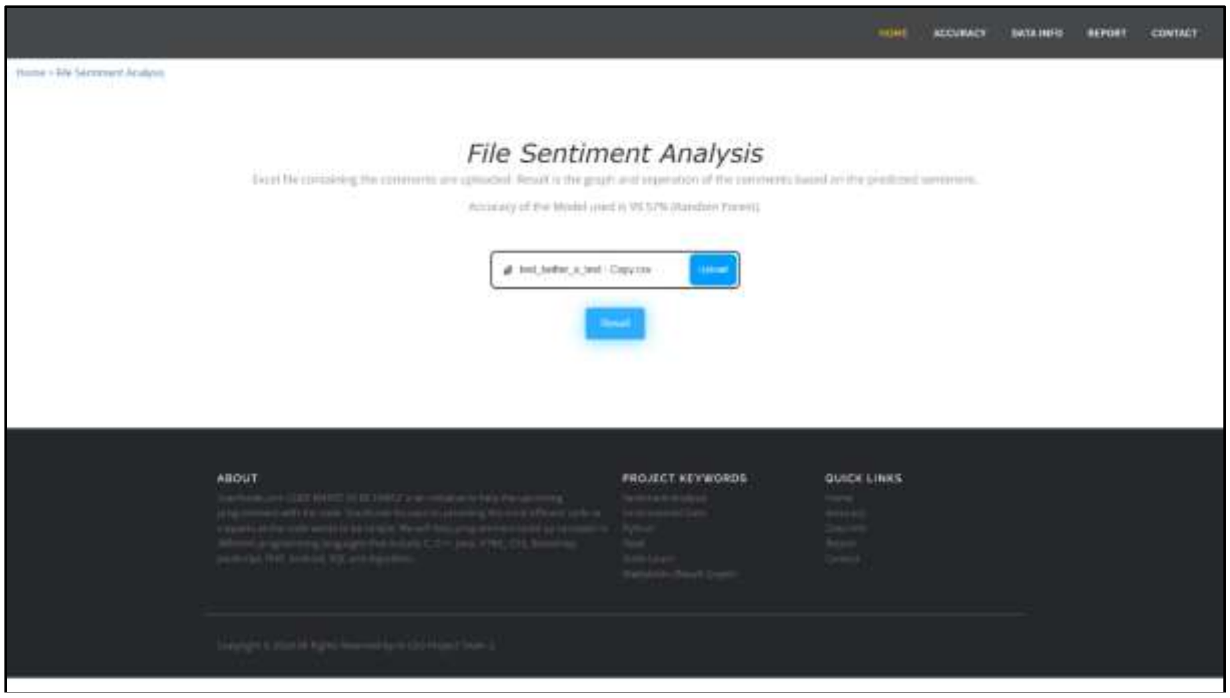
9. Result

Our project website serves as an intuitive platform for users to effortlessly input their data and receive accurate predictions regarding house prices. With a user-friendly interface, the site seamlessly guides users through the process of uploading files, enabling them to obtain results presented through pie charts and bar graphs. Additionally, the website provides insightful comments on the sentiment associated with each prediction, enhancing the user experience and facilitating informed decision-making


Sentiment Analysis

Process the tweets and gives the output as Positive, Negative and Neutral

Fig.3. Proposed System Home Page



File Sentiment Analysis

Excel file containing the comments are uploaded. Result is the graph and separation of the comments based on the predicted sentiment.

Accuracy of the Model used is 95.57% (Random Forest)

test_twitter_x_test - Copy.csv

ABOUT

[illegible]

PROJECT KEYWORDS

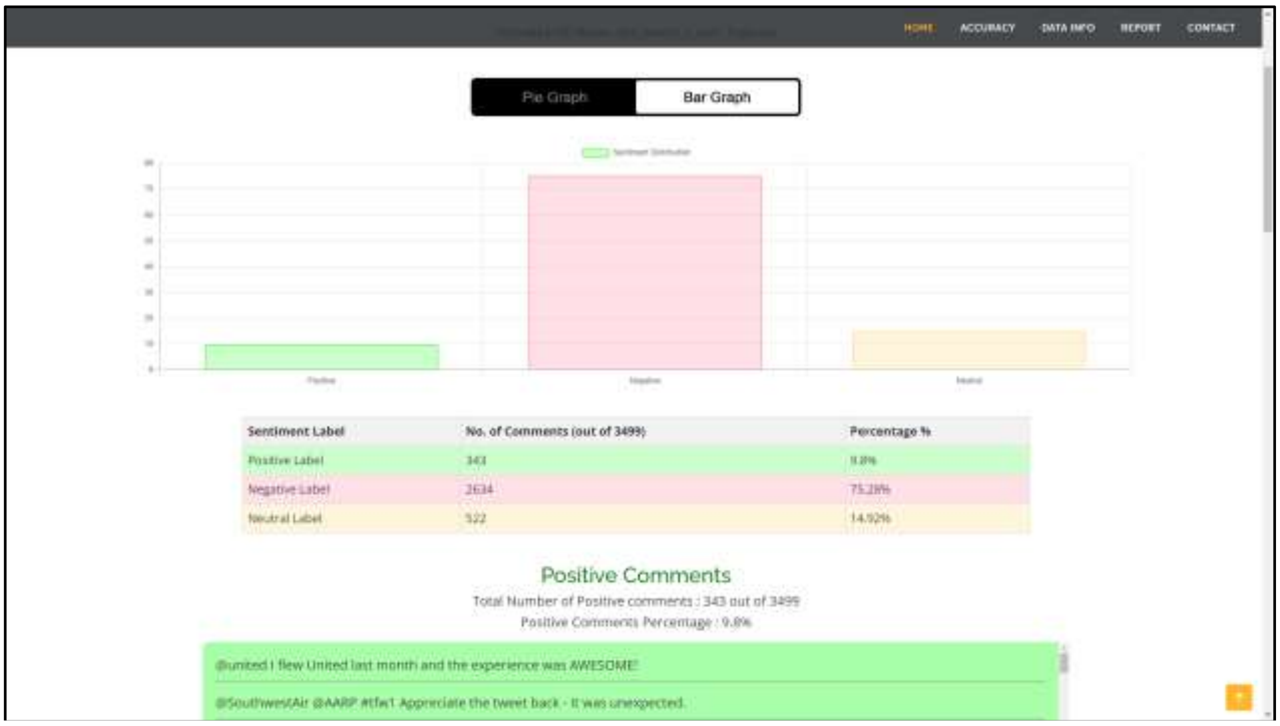
Sentiment Analysis
[illegible]
Python
Flask
Scikit Learn
Matplotlib (Result Graph)

QUICK LINKS

Home
Accuracy
Data Info
Report
Contact

Copyright © 2024 All Rights Reserved by [illegible]

Fig.4. File Upload Page



Sentiment Label	No. of Comments (out of 3499)	Percentage %
Positive Label	343	9.8%
Negative Label	2634	75.28%
Neutral Label	522	14.92%

Positive Comments

Total Number of Positive comments : 343 out of 3499

Positive Comments Percentage : 9.8%

@united I flew United last month and the experience was AWESOME!

@SouthwestAir @AARP #tfw1 Appreciate the tweet back - It was unexpected.

Fig.5. Result Page

10. Future Scope

The future of sentiment analysis is likely to involve advancements in several key areas:

- **Deep Learning and AI:** Further improvements in deep learning models and artificial intelligence techniques will lead to more accurate sentiment analysis systems. This may involve more sophisticated architectures such as transformers, which have shown significant promise in natural language processing tasks.
- **Multimodal Sentiment Analysis:** Integrating text with other modalities like images, audio, and video will enable more comprehensive sentiment analysis. This will be particularly useful for applications like social media monitoring, where content is often multimodal.
- **Contextual Understanding:** Enhancing sentiment analysis models to better understand context will be crucial. This includes understanding sarcasm, irony, and other forms of nuanced language, as well as considering the broader context in which a statement is made.
- **Domain-Specific Analysis:** Customizing sentiment analysis models for specific domains, such as finance, healthcare, or customer service, will lead to more accurate results tailored to the unique language and expressions used in those domains.

- **Cross-Lingual Sentiment Analysis:** Developing models that can analyze sentiment in multiple languages will be important for global applications. This will involve overcoming challenges such as language differences, cultural nuances, and the scarcity of labeled data for certain languages.
- **Ethical Considerations:** As sentiment analysis becomes more widespread, ethical considerations surrounding privacy, bias, and fairness will become increasingly important. Ensuring that sentiment analysis systems are used responsibly and do not perpetuate or amplify biases will be a significant focus.
- **Real-Time Analysis:** Improving the speed and scalability of sentiment analysis systems will enable real-time monitoring of sentiment on social media, news articles, and other sources of textual data. This will be valuable for businesses and organizations to stay informed about public opinion as it evolves.

Conclusion

In conclusion, sentiment analysis on Twitter data presents a crucial avenue for understanding public opinion and sentiment trends in real-time. Our project offers a robust framework leveraging machine learning algorithms to categorize tweets accurately. By emphasizing feature extraction, data preprocessing, and model evaluation, we streamline the sentiment analysis process for actionable insights. Looking forward, advancements in deep learning, multimodal analysis, and ethical considerations will drive further innovation in the field. Sentiment analysis continues to play a pivotal role in informing decision-making processes and understanding human behavior, contributing to a more connected and informed society.

References

[1] Sentiment Analysis of Twitter Data: A Survey of Techniques: Vishal A. Kharde & S.S. Sonawane, 11, April 2016.

[2] A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326.

[3] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011Workshop on Languages in Social Media,2011 , pp. 30-38.

[4] Neethu M,S and Rajashree R," Sentiment Analysis in Twitter using Machine Learning Techniques" 4th ICCCNT 2013,at Tiruchengode, India. IEEE – 31661.

[5] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[6] Go, R. Bhayani, L.Huang. "Twitter Sentiment ClassificationUsing Distant Supervision". Stanford University, Technical Paper,2009

[7] Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data", In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

[8] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, "Using wordnet to measure semantic orientations of adjectives," 2004.