

International Journal of
Engineering Research and Science & Technology



ISSN : 2319-5991

www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

TWITTER SENTIMENT ANALYSIS USING MACHINE LEARNING

¹ Dr.A. Avani, ² T. Bharath Krishna

¹ Associate Professor, Department of Computer Science and Engineering

Anubose Institute of Technology, New Palvoncha-507115,

Bhadradi Kothagudem-Dist-TG

² Department of Computer Science and Engineering

Anubose Institute of Technology, New Palvoncha-507115,

Bhadradi Kothagudem-Dist-TG

ABSTRACT

Twitter is a platform widely used by people to express their opinions and display sentiments on different occasions. Sentiment analysis is an approach to analyze data and retrieve sentiment that it embodies. Twitter sentiment analysis is an application of sentiment analysis on data from Twitter (tweets), in order to extract sentiments conveyed by the user. In the past decades, the research in this field has consistently grown. The reason behind this is the challenging format of the tweets which makes the processing difficult. The tweet format is very small which generates a whole new dimension of problems like use of slang, abbreviations etc. We review some papers regarding research in sentiment analysis on Twitter, describing the methodologies adopted and models applied, along with describing a generalized Python based approach. So, this project aims to build a system that identifies the human emotions and opinions expressed in text. This system uses machine learning models to collect the tweets-related data, classify them as positive, negative, or neutral, and evaluate the results.

1.INTRODUCTION

Now-a-days social networking sites are at the boom, so large amount of data is generated. Millions of people are sharing their views daily on micro blogging sites, since they contain short and simple expressions. We shall discuss about a paradigm to extract the sentiment from a famous micro blogging service, Twitter, where users tweet their opinions about different things. In this project, we will discuss the sentiment analysis of twitter dataset with data mining

approach. An approach is introduced that automatically classifies the sentiments of Tweets taken from Twitter dataset. These messages or tweets are classified as positive, negative or neutral. In this procedure of sentiment analysis, we will take the Twitter dataset as input and train the system to understand the positive, neutral and negative tweets and cluster them as individual data chunks. The training data consists of tweets which can be plain-text, acronyms,

emoticons and abbreviations. This is very useful for the companies who want to know the feedback about their product brands and customer purchase product details which in turn is helpful for them to develop and gain profits. Despite advancements in sentiment analysis techniques, there are still significant challenges. One challenge is the handling of context and ambiguity in language. For instance, a tweet like “I love this product, but the service is terrible” expresses mixed sentiments that need to be separated for a meaningful analysis. Furthermore, the evolving nature of language, with the introduction of new slang and acronyms, requires continuous updates to the models and lexicons used for sentiment classification. The domain of a tweet also influences sentiment analysis—tweets about entertainment, politics, and technology, for example, may use different vocabulary and expressions that require domain-specific models. The applications of Twitter sentiment analysis are vast. In the business world, companies monitor Twitter to detect customer satisfaction, identify potential problems, and track brand health. In politics, sentiment analysis can help understand voter behaviour, the popularity of political leaders, or the public's opinion on policy changes. During crises, such as natural disasters or pandemics, sentiment analysis can also provide real-time insights into public emotions and reactions, helping organizations respond effectively. Twitter sentiment analysis refers to the process of using natural language processing (NLP) and machine learning techniques to determine the emotional tone behind Twitter posts or tweets. By analyzing the language used in tweets, sentiment analysis aims to classify them as positive, negative, or neutral,

providing insights into public opinion on a specific topic, event, brand, or individual. Twitter, with its vast user base and real-time nature, makes it an ideal platform for sentiment analysis. Tweets often express spontaneous reactions, opinions, and feelings, which can help gauge public sentiment on various subjects. The analysis can be performed at multiple levels, including individual tweets, trends, or hashtags. Sentiment analysis on Twitter has various applications. In business, it helps brands monitor customer satisfaction and track the effectiveness of marketing campaigns. In politics, it provides insights into voter sentiment, while in crisis management, it helps track public reactions to ongoing events. However, challenges remain, including the handling of sarcasm, slang, and context-specific expressions, which can complicate accurate sentiment classification. Despite these challenges, Twitter sentiment analysis continues to be a valuable tool for understanding public sentiment in an increasingly connected and fast-paced world.

II.LITERATURE SURVEY

2.1SENTIMENT ANALYSIS

Sentiment Analysis is process of collecting and analyzing data based upon the person feelings, reviews and thoughts. Sentimental analysis often called as opinion mining as it mines the important feature from people opinions. Sentimental Analysis is done by using various machine learning techniques, statistical models and Natural Language Processing (NLP) for the extraction of feature from a large data. Sentiment Analysis can be done at document, phrase and sentence level. In document level, summary of the

entire document is taken first and then it is analyze whether the sentiment is positive, negative or neutral. In phrase level, analysis of phrases in a sentence is taken in account to check the polarity. In Sentence level, each sentence is classified in a particular class to provide the sentiment. Sentimental Analysis has various applications. It is used to generate opinions for people of social media by analyzing their feelings or thoughts which they provide in form of text. Sentiment Analysis is domain centered, i.e. results of one domain cannot be applied to other domain. Sentimental Analysis is used in many real life scenarios, to get reviews about any product or movies, to get the financial report of any company, for predictions or marketing. Twitter is a micro blogging platform where anyone can read or write short form of message which is called tweets. The amount of data accumulated on twitter is very huge. This data is unstructured and written in natural language. Twitter Sentimental Analysis is the process of accessing tweets for a particular topic and predicts the sentiment of these tweets as positive or negative with the help of different machine learning algorithms.

1. With the advancement of web technology and its growth, there is a huge volume of data present in the web for internet users and a lot of data is generated too.

2. Internet has become a platform for online learning, exchanging ideas and sharing opinions. Social networking sites like Twitter, Facebook, Google+ are rapidly gaining

3. popularity as they allow people to share and express their views about topics, have

discussion with different communities, or post messages across the world.

4. There has been lot of work in the field of sentiment analysis of twitter data. This project focuses mainly on sentiment analysis of twitter data which is helpful to analyse the

information in the tweets where opinions are highly unstructured, heterogeneous and are either positive or negative, or neutral in some cases.

5. Sentiment analysis can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through Natural Language Processing (NLP).

6. Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". It is also referred as subjectivity analysis, opinion mining, and appraisal extraction.

7. The words opinion, sentiment, view and belief are used interchangeably but there are differences between them.

Opinion: A conclusion opens to dispute (because different experts have different opinions)

View: subjective opinion

Belief: deliberate acceptance and intellectual assent

Sentiment: opinion representing one's feelings

2.2 PYTHON

Python is a high level, dynamic programming language which is used for this thesis. Python 3.4 version was used as it is a mature, versatile and robust programming

version. It is an interpreted language which makes the testing and debugging extremely quickly as there is no compilation step. There are extensive open source libraries available for this version of python and a large community of users. Python is simple yet powerful, interpreted and dynamic programming language, which is well known for its functionality of processing natural language data, i.e. spoken English using NLTK. Other high level programming languages such as 'R' and 'Mat lab' were considered because they have many benefits such as ease of use but they do not offer the same flexibility and freedom that Python can deliver. Python (Python 3.4) language is used as a programming platform for this twitter sentiment analysis.

1. Python is chosen because it is the most popular open-source application with huge community support and it is best for engineers and programmers. Whereas R-tool is best for scientists and researchers.

2. Moreover Python provides us with Natural Language Toolkit (NLTK) which makes our coding lot more easier on sentiment analysis.

3. Python can also be used for scripting a website to display the results on web pages.

4. Python is supported by vast number of modules and we can import any module we want which in turn reduces our tedious work.

5. Scipy, Matplotlib and Pygame modules are used so as to generate graphs and calculations.

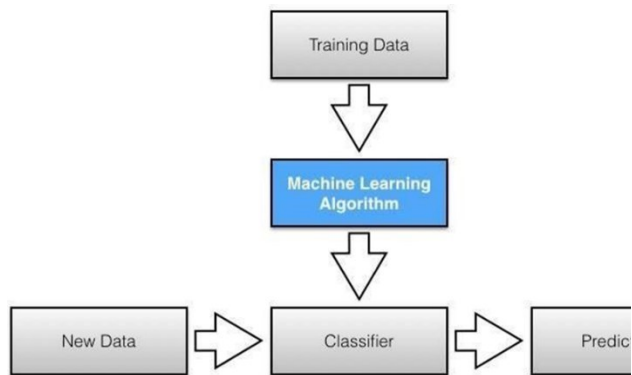
2.3 NLTK

Natural Language Toolkit (NLTK) is library in Python, which provides a base for building programs and classification of data. NLTK is

a collection of resources for Python that can be used for text processing, classification, tagging and tokenization. This toolbox plays a key role in transforming the text data in the tweets into a format that can be used to extract sentiment from them. NLTK provides various functions which are used in pre-processing of data so that data available from twitter become fit for mining and extracting features. NLTK support various machine learning algorithms which are used for training classifier and to calculate the accuracy of different classifier. In our thesis we use Python as our base programming language which is used for writing code snippets. NLTK is a library of Python which plays a very important role in converting natural language text to a sentiment either positive or negative. NLTK also provides different sets of data which are used for training classifiers. These datasets are structured and stored in library of NLTK, which can be accessed easily with the help of Python.

III. SYSTEM ARCHITECTURE :

We use different feature sets and machine learning classifiers to determine the best combination for sentiment analysis of twitter. We also experiment with various pre-processing steps like - punctuations, emoticons, twitter specific terms and stemming. We investigated the following features - unigrams, bigrams, trigrams and negation detection. We finally train our classifier using various machine-learning algorithms - Naive Bayes, Decision Trees and Maximum Entropy. We present a new feature vector for classifying the tweets as positive, negative and extract peoples' opinion about products.



possible deployment options or settings. The application's clean and functional design is geared toward testing or development purposes. The desktop environment in the background further indicates that the application might still be in a testing phase or locally hosted for initial evaluations.

It represents a straightforward sentiment analysis tool.

Negative Interface:

IV.OUTPUT SCREENS

INTERFACE:

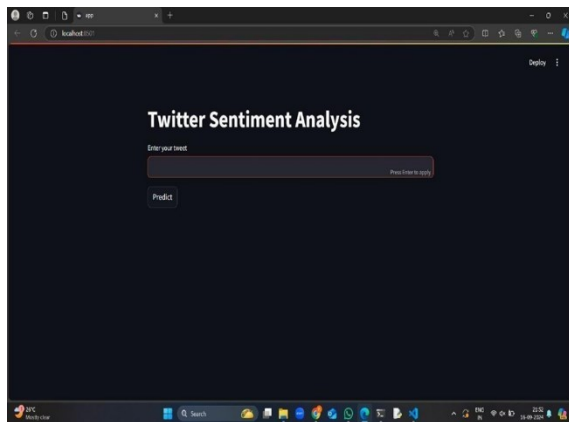


Fig 4.1

The image showcases a Twitter Sentiment Analysis application with a simple and intuitive user interface. It runs on a local server, as indicated by the localhost:8501 address in the browser's address bar. The interface has a dark theme and features a text input box where users can type in a tweet.

Below this, a "Predict" button initiates the sentiment analysis process

The input field is currently empty, and a placeholder text, "Please Enter to apply," is visible, suggesting the user must press "Enter" after typing a tweet. The top-right corner displays a "Deploy" menu, hinting at

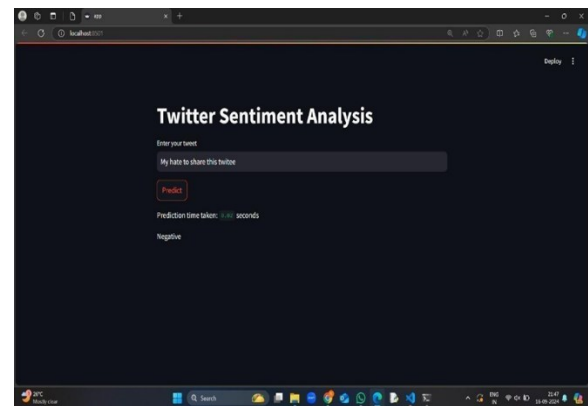


Fig 4.2

The image displays a web-based application for Twitter Sentiment Analysis. The interface is minimalistic, with a text input box where users can type or paste a tweet. Below the input field, there is a "Predict" button that triggers the sentiment analysis process. The sample tweet entered in the box reads, "My hate to share this tweet." Once the prediction is performed, the app displays the sentiment as "Negative," with the prediction time noted as 0.02 seconds, emphasizing the tool's efficiency. The page uses a dark theme, making it visually appealing and easier on the eyes. The design is straightforward, catering to ease of use for end-users. At the top-right corner, a menu button labelled "Deploy" suggests additional functionalities or

deployment options might be available. The image also includes a glimpse of a desktop environment, showcasing the application running on localhost. This appears to be a project likely under testing or development.

Positive Interface:

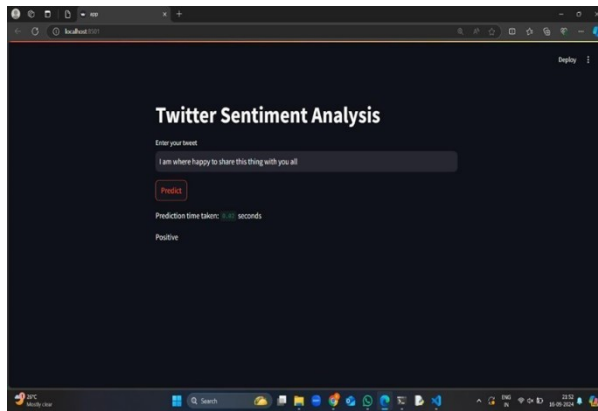


Fig 4.3

This image represents the output screen of a Twitter Sentiment Analysis web application with a positive sentiment prediction. The interface features a dark theme and a simple, user-friendly layout. At the top, the title "Twitter Sentiment Analysis" is displayed prominently. Below it, there is an input field labeled "Enter your tweet," where users can type their text.

For a positive output, the input could be a statement like, "i am where happy to share this thing with you all!" Once the user clicks the "Predict" button, the application quickly processes the text and displays the result. The prediction output would indicate "Positive," along with a short prediction time, such as 0.02 seconds.

This clear and efficient interface demonstrates the functionality of the application, providing accurate and fast sentiment predictions. It is ideal for analyzing

public opinions or emotions on social media posts in real time.

Neutral Interface:

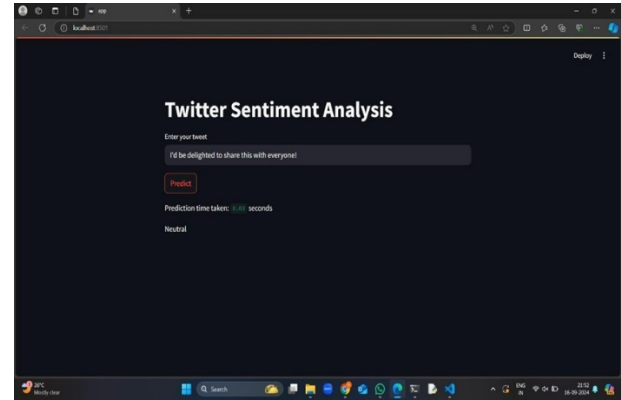


Fig 4.4

This image shows the output screen of a Twitter Sentiment Analysis web application, featuring a neutral sentiment prediction. The interface maintains a clean, dark-themed design with a straightforward structure. At the top, the title "Twitter Sentiment Analysis" is prominently displayed. Below it, an input field labelled "Enter your tweet" allows users to type their text.

For a neutral output, the input might be a statement like, "I'd be delighted to share this with everyone." After the user clicks the "Predict" button, the application processes the text within seconds and displays the result. In this case, the prediction would read "Neutral," along with the processing time, such as 0.02 seconds.

This tool offers a user-friendly way to analyze tweets and determine whether the sentiment is positive, negative, or neutral. It provides valuable insights into public opinion or general

V.CONCLUSION

We conclude by saying that, our project can be used by any company or industry to conduct a survey about their products so as to determine the statistics of their business. Customer analysis can be done to improvise their business. Can be applicable to movie Review-related websites such as movie reviews, product reviews etc. Individual user can also use it to know the sentiment about a particular product, topic or politics. The usage of latest and updated NLTK classifiers improve our accuracy and more over we improve the training set to get more accurate results. Sentiment analysis of Twitter data has become an important task in machine learning, given the massive amount of data generated daily on the platform. Twitter is a rich source of user opinions, making it invaluable for businesses, governments, and researchers looking to understand public sentiment on a variety of topics. The goal of sentiment analysis is to automatically determine whether a tweet expresses a positive, negative, or neutral opinion. This type of analysis can provide insights into consumer preferences, political trends, or public perception of events, products, and services. Over the years, advancements in natural language processing (NLP) and machine learning (ML) techniques have made it possible to analyze this unstructured data with increasing accuracy. One of the primary challenges in sentiment analysis is the complexity and variability of human language. Tweets often contain slang, abbreviations, emojis, hashtags, and other forms of informal communication that make the task of sentiment classification challenging. Additionally, sarcasm, irony, and context-dependent meanings can lead to misinterpretation by machine learning models. These challenges necessitate the

development of sophisticated methods to accurately detect sentiment. Traditional machine learning techniques for sentiment analysis rely heavily on feature extraction from the text data. Methods like bag-of-words (BoW) or term frequency-inverse document frequency (TF-IDF) are often used to transform the text into a numerical form that machine learning algorithms can understand. However, these methods can be limited in capturing the semantic meaning of words. More advanced approaches, such as word embeddings (Word2Vec, GloVe), have been introduced to overcome these limitations. Word embeddings map words to dense vector representations that capture their semantic relationships, allowing for a better understanding of word meanings in context.

Another breakthrough in sentiment analysis has been the rise of transformer models, particularly BERT (Bidirectional Encoder Representations from Transformers) and its variants. BERT has been shown to outperform previous models on a variety of NLP tasks, including sentiment analysis. Unlike traditional methods that process text from left to right or right to left, BERT takes a bidirectional approach, capturing context from both directions simultaneously. This enables the model to better understand the relationships between words and their meanings, especially in complex and nuanced sentences. Pre-trained models such as BERT can be fine-tuned on specific sentiment analysis datasets, enabling them to adapt to the particular domain of Twitter data.

While deep learning techniques have provided significant improvements in sentiment analysis, they require large amounts of labeled data for training, which is

often a bottleneck in many applications. To address this challenge, several approaches such as transfer learning and semisupervised learning have been developed. Transfer learning allows models pre-trained on large corpora of text data to be fine-tuned on smaller, domain-specific datasets. This approach has been particularly useful for sentiment analysis tasks, where labeled data may be scarce. Semi-supervised learning combines both labeled and unlabeled data to train models more effectively, using techniques like self-training or co-training to iteratively improve the model's performance. Another consideration in sentiment analysis of Twitter data is the balance between model performance and interpretability. While deep learning models tend to perform better on sentiment classification tasks, they are often considered "black boxes," meaning it is difficult to understand why they make certain predictions. In many applications, such as in business or politics, it is crucial to not only predict sentiment accurately but also to explain why a certain prediction was made. Efforts have been made to improve the interpretability of machine learning models, such as using attention mechanisms in neural networks to highlight which words or phrases in a tweet influenced the model's decision. This can help to provide transparency and increase trust in the model's predictions. The data used for sentiment analysis in Twitter comes in many forms, from simple text tweets to those enriched with hashtags, mentions, and multimedia content. However, the challenge is that tweets are often short, filled with slang, abbreviations, emojis, and other non-standard forms of expression, making it harder for traditional machine learning models to correctly infer sentiment. To address this, researchers often preprocess

the text by removing noise (like URLs, stop words, and special characters) while also expanding abbreviations and handling negations, all of which can significantly affect the sentiment of a tweet. The preprocessing steps are essential in ensuring that the features used for model training are relevant and contribute to the accuracy of predictions. When it comes to evaluating sentiment, the typical categories are positive, negative, and neutral, though some systems include more granular classifications, such as mixed sentiment or emotions like anger, joy, or sadness. One of the key challenges in sentiment analysis is the ambiguity of language. A single tweet might contain contradictory sentiments or complex expressions that are difficult to categorize. For instance, a sarcastic tweet might appear positive at first glance but could be negative upon closer inspection, which makes it challenging to train models that generalize well across different types of tweets. This is why understanding the broader context of the conversation and the user's history is important, but it also introduces new challenges for machine learning models in terms of capturing that context.

VI. FUTURE SCOPE

We look forward to use bigger dataset to improve the accuracy

We look forward to consider emoticons, expressions and internet slangs in our training data

We look forward to classify tweets as neutral tweets along with positive and negative tweets

The scope for further advancements in Twitter sentiment analysis using machine

learning is broad, and it holds significant potential to transform how we understand public opinions and emotions expressed online. As social media platforms continue to grow, particularly Twitter, the amount of data generated becomes increasingly complex, and the need for effective sentiment analysis models grows with it. While existing models are capable of processing basic sentiment classification tasks, there are several areas where future research and advancements could significantly enhance the effectiveness and applicability of sentiment analysis. A key area for future development is improving the ability of machine learning models to understand deeper contextual nuances within tweets. While current models, such as deep learning-based approaches like LSTMs and transformers, have demonstrated significant advancements, understanding sarcasm, irony, humor, and subtle shifts in sentiment remains a challenge. Many tweets convey sentiment indirectly, through irony or sarcasm, which requires a more sophisticated level of interpretation that current models often fail to capture. For example, a tweet that reads "Great, another Monday!" is sarcastic, expressing negative sentiment despite the positive wording. Models that can effectively understand and differentiate between direct and indirect sentiment will be crucial in improving accuracy. This could involve advanced models that can detect emotional undertones and subtler forms of sentiment that go beyond a simple positive or negative classification.

Another important area for growth is the incorporation of multimodal sentiment analysis. Tweets on Twitter often contain more than just text—users post images, videos, emojis, and use hashtags, all of which can convey meaning. Current sentiment

analysis models that only consider textual data miss the sentiment conveyed through these additional media. For example, a tweet containing a positive text paired with a happy emoji can indicate a stronger sentiment than the text alone. Similarly, the meaning of hashtags like #love or #hate can shift the overall sentiment of the tweet. By developing models that integrate multiple forms of media, machine learning can offer a more comprehensive and nuanced understanding of sentiment in Twitter posts. Multimodal models that combine text with images or emojis can enhance the accuracy of sentiment analysis by capturing the full context of a tweet. Additionally, there is a need for more fine-grained sentiment analysis. Current models typically classify tweets as positive, negative, or neutral, but this broad classification doesn't always capture the complexity of sentiment in tweets. For instance, tweets can express varying degrees of positivity or negativity, or they can convey mixed emotions about a subject. A tweet like "I really enjoy the features of this phone, but the battery life is disappointing" expresses both positive and negative sentiment. Fine-grained sentiment analysis aims to not only detect whether a sentiment is positive or negative but also to measure the intensity of that sentiment and to identify different emotional aspects such as happiness, anger, or sadness. The ability to capture these subtleties can provide more detailed insights into public opinion, allowing businesses and organizations to make more informed decisions. Another promising area for future research is the integration of multimodal data in sentiment analysis. While Twitter is primarily a text-based platform, many tweets contain images, videos, and even audio content, which may carry important context

or sentiment that is not captured by text alone. By incorporating computer vision and audio processing into sentiment analysis, it may be possible to develop more accurate models that understand the full context of a tweet, including visual cues, tone of voice, and non-verbal expressions. Multimodal sentiment analysis could enable a deeper understanding of social media content and offer more nuanced insights into public sentiment, particularly in cases where the text itself is ambiguous or lacking in emotional cues. Sentiment analysis on Twitter can also benefit from advancements in fine-grained sentiment detection. Traditional sentiment analysis often classifies tweets as simply positive, negative, or neutral. However, there are cases where sentiments are more complex and involve mixed emotions, such as a tweet expressing both frustration and hope or excitement and anxiety. Developing models that can classify more granular levels of sentiment or detect mixed sentiments would open up new opportunities for understanding public opinion in greater detail. This could be especially useful in areas like brand sentiment analysis, political discourse, and crisis management, where a single sentiment category may not fully capture the complexity of the audience's feelings.

VII. REFERENCES

- 1.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
- 2.R. Parikh and M. Movassate, "Sentiment Analysis of User- Generated Twitter Updates using Various Classi_cation Techniques",CS224N Final Report, 2009
- 3.Go, R. Bhayani, L.Huang. "Twitter Sentiment Classification Using Distant Supervision". Stanford University, Technical Paper,2009
- 4.L. Barbosa, J. Feng. "Robust Sentiment Detection on Twitter from Biased and Noisy Data". COLING 2010: Poster Volume,pp. 36-44.
- 5.Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", In Proceedings of the 13th International Conference on Discovery Science, Berlin, Germany: Springer,2010, pp. 15.
- 6.Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011 Workshop on Languages in Social Media,2011 , pp. 3038
- 7.Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volume pages 241{249, Beijing, August 2010
- 8.Po-Wei Liang, Bi-Ru Dai, "Opinion Mining on Social Media Data", IEEE 14th International Conference on Mobile Data Management,Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 97814946736068-5, <http://doi.ieeecomputersociety.org/10.1109/MDM.2013>.
- 9.Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets", 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-175.