IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Analyzing Event Hype and Reality on Twitter

A Sentiment Analysis Approach using Tweepy and Preprocessing Techniques

[1]Ms. Arthi Basta, [2]Sandeep Poloju, [3]Sai Sampath [4]SaiTeja Reddy

[1] Assistant Professor, Department of Computer science and Engineering, CVR College of Engineering, Hyderabad, Telangana, India

[2,3,4] IV BTech Students, Department of Computer science and Engineering, CVR College of Engineering, Hyderabad, Telangana, India

ABSTRACT

Social media websites have emerged as one of the platforms to raise users' opinions and influence the way any business is commercialized. Opinion of people matters a lot to analyze how the propagation of information impacts the lives in a large-scale network like Twitter. This research paper focuses on sentiment analysis of Twitter data to evaluate the level of hype and reality associated with specific events. Twitter data related to various events, categorized as pre-event, during-event, and post-event, is analysed to determine the sentiment polarity and user inclination. The research contributes by demonstrating the effectiveness and applicability of these sentiment analysis methods for event analysis on Twitter. The study employs established sentiment analysis techniques, namely Naive Bayes and Support Vector Machines (SVM), combined with a comprehensive set of pre-processing techniques.

Keywords: Sentiment analysis, Naive Bayes, Support Vector Machines.

1. INTRODUCTION

- Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials, Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document
- The attitude may be his or her judgment or evaluation affective state, or the intended emotional communication, Sentiment analysis is the process of detecting a piece of writing for positive, negative, or neutral feelings bound to it
- Sentiment analysis plays a crucial role in comprehending and analyzing the sentiment, attitudes, and emotions expressed by users on social media.
- The objective of this research paper is to analyze the sentiment of Twitter data specifically related to events and evaluate the level of hype and reality associated with these events.
- By conducting sentiment analysis on event-related tweets, we aim to provide insights into public sentiment, perception, and the potential discrepancy between the hype generated and the actual reality of events.
- This research seeks to contribute to the understanding of event analysis on social media platforms, particularly Twitter, and its implications for event management and decision-making processes.

2. LITERATURE SURVEY

Efthymios Kouloumpis, Theresa Wilson, Johns Hopkins University, USA, Johanna Moore, School of Informatics University of Edinburgh, Edinburgh, UK in a paper on **Twitter Sentiment Analysis:The Good the Bad and the OMG! in July 2011** have investigate the utility of linguistic features for detecting the sentiment of Twitter messages. We evaluate the usefulness of existing lexical resources as well as features capture information about the informal and creative language used in microblogging. We take a supervised approach to the problem, but leverage existing hashtags in the Twitter data for building training data.

Hassan Saif, Yulan He and Harith Alani, Knowledge Media Institute, The Open University, United Kingdom in a paper **Semantic Sentiment Analysis of Twitter in Nov 2012** they have introduce a novel approach of adding semantics as additional features into the training set for sentiment analysis. For each extracted entity (e.g. iPhone) from tweets, we add its semantic concept (e.g. "Apple product") as an additional feature, and measure the correlation of the representative concept with negative/positive sentiment.

Subhabrata Mukherjee1, Akshat Malul, Balamurali A.R.12, Pushpak Bhattacharyya1,1Dept. of Computer Science and Engineering, IIT Bombay, 211TB- Monash Research Academy, IIT Bombay on a paper on **TwiSent: A Multistage System for Analyzing Sentiment in Twitter in Feb 2013** they have presented TwiSent, a sentiment analysis system for Twitter. Based on the topic searched, TwiSent collects tweets pertaining to it and categorizes them into the different polarity classes positive, negative and objective. However, analyzing micro-blog posts have many inherent challenges compared to the other text genres.

Isaac G. Councill, Ryan McDonald, Leonid Velikovich, Google, Inc., New York on a paper on **What's Great and What's Not: Learning to Classify the Scope of Negation for Improved Sentiment Analysis in July 2010** presents a negation detection system based on a conditional random field modelled using features from an English dependency parser. The scope of negation detection is limited to explicit rather than implied negations within single sentences.

Dr. R. Usha Rani, Arthi Basta presents **Twitter Sentiment and Statement Reality Analysis** to evaluate the PreCurrent and Post Ranking analysis will give Reality of the Tweet. The author counts the individual sentiment to achieve the reality check.

3. EXISTING APPROACHES TO SENTIMENT ANALYSIS

Existing approaches to sentiment analysis can be grouped into three main categories:

- Keyword spotting
- Lexical affinity
- Statistical methods

Keyword spotting is the most naive approach and probably also the most popular because of its accessibility and economy .Text is classified into affect categories based on the presence of fairly unambiguous affect words like 'happy', 'sad', 'afraid', and 'bored" .The weaknesses of this approach lie in two areas: poor recognition of affect when negation is involved and reliance on surface features About its first weakness, while the approach can correctly classify the sentence "today was a happy day" as being happy, it is likely to fail on a sentence like "today wasn't a happy day at all" About its second weakness, the approach relies on the presence of obvious affect words that are only surface features of the prose. In practice, a lot of sentences convey affect through underlying meaning rather than affect adjectives For example, the text "My husband just filed for divorce and he wants to take custody of my children away from me" certainly evokes strong emotions, but uses no affect keywords, and therefore, cannot be classified using a keyword spotting approach.

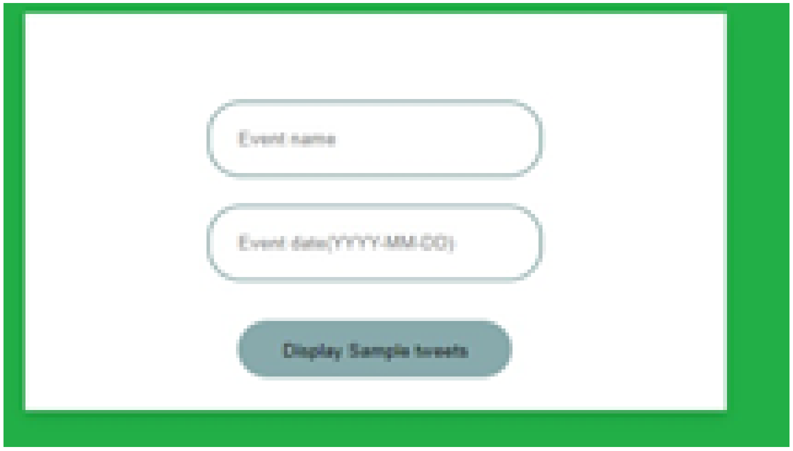
Lexical affinity is slightly more sophisticated than keyword spotting as, rather than simply detecting obvious affect words, it assigns arbitrary words a probabilistic affinity for a particular emotion For example, "accident might be assigned a 75% probability of being indicating a negative affect, as in car accident' or 'hurt by accident These probabilities are usually trained from linguistic corpora. Though often outperforming pure keyword spotting.

there are two main problems with the approach First, lexical affinity, operating solely on the word-level, can easily be tricked by sentences like "I avoided an accident" (negation) and "I met my girlfriend by accident" (other word senses) Second, lexical affinity probabilities are often biased toward text of a particular genre, dictated by the source of the linguistic corpora this makes it difficult to develop a reusable, domain-independent model.

Statistical methods, such as Bayesian inference and support vector machines, have been popular for affect classification of texts By feeding a machine learning algorithm a large of affectively annotated texts, it is possible for the system to not only learn the training corpus of , affective valence of affect keywords (as in the keyword spotting approach), but also to take into account the valence f other arbitrary keywords (like lexical affinity), punctuation, and word co-occurrence frequencies. However, traditional statistical methods are generally semantically weak, meaning that, with the exception of obvious affect keywords, other lexical or co-occurrence elements in a statistical model have little predictive value individually. As a result, statistical text classifiers only work with acceptable accuracy when given a sufficiently large text input.So, while these methods may be able to affectively classify user's text on the page or paragraph-level,they do not work well on smaller text units such as sentences or clauses.

4. METHODOLOGY

- The methodology employed in this research involves collecting tweets related to specific events using the tweepy library, which accesses the Twitter API and allows retrieval based on event names or hashtags.
- The collected tweets are then categorized into pre-event, during-event, and post-event categories based on their timestamps.
- A comprehensive set of pre-processing techniques is applied to clean and normalize the collected tweets. This includes removing usernames, URLs, punctuation, and stop words, as well as expanding acronyms and applying stemming and lemmatization.
- In addition to the collected tweets, the sentiment140 dataset from Kaggle, consisting of 1.6 million labelled tweets, is utilized as the primary data source for training and testing the sentiment analysis models.
- The sentiment analysis models, such as Naive Bayes and Support Vector Machines (SVM), are trained on the sentiment140 dataset to learn patterns and sentiments associated with different tweets.
- The pre-processed collected tweets are then used as test data for the trained sentiment analysis models to evaluate their performance and assess sentiment polarity.
- The sentiment analysis results are used to determine the sentiment polarity and inclination of users towards the specific events.
- The effectiveness and applicability of the sentiment analysis techniques, combined with the pre-processing steps and tweet categorization, are evaluated in the context of event analysis on Twitter.
- The research aims to assess the level of hype and reality associated with the events by analyzing the sentiment of tweets before, on, and after the event, using both the collected tweets and the sentiment140 dataset.
- By incorporating the sentiment140 dataset, the research also provides insights into the generalizability and performance of the sentiment analysis models on a large-scale dataset.
- Ultimately, interface takes the event name and event date and provides the results



5. RESULTS

Here we are tabulating the results with parameters for better understanding of what model to choose with all the parameter tuning.

Naive Bayes

	Unigram	Unigram and Bigram
few_preprocessing_steps	0.7623	0.7746
After stemming	0.7731	0.7843
After lemmetizing	0.7731	0.7843

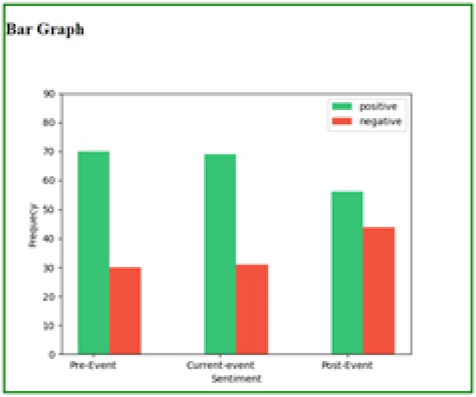
naïve bayes results table

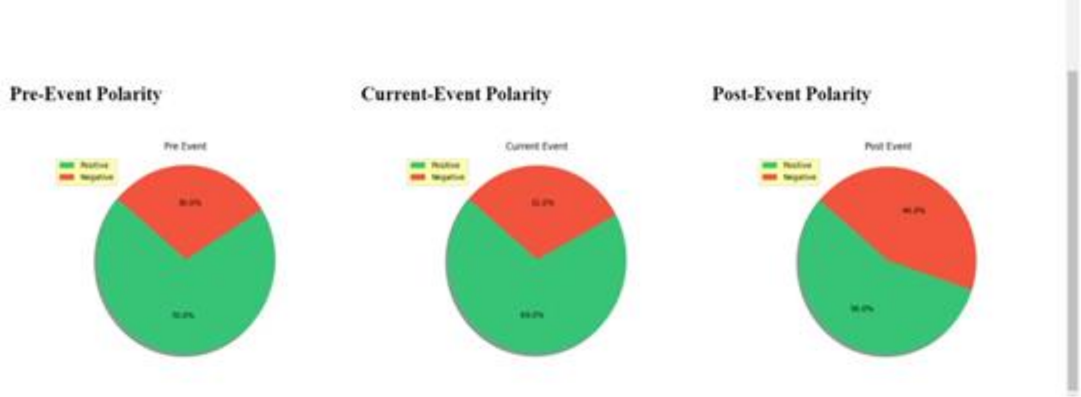
SVM

	Unigram	Unigram and Bigram
few_preprocessing_steps	0.7403	0.7501
After stemming	0.7403	0.7562
After lemmetizing	0.7477	0.7562

SVM results table

Results of polarity can be evaluated


Bar Graph


Pre-Event Polarity

Current-Event Polarity

Post-Event Polarity

6. CONCLUSION & FUTURE SCOPE

Sentiment Analysis is the process of detecting positive or negative sentiment in text. It’s often used by businesses to detect sentiment in social data, gauge brand reputation, and understand customers.

In this project, we have gone through multiple preprocessing stages for large dataset.We have performed Sentiment analysis of the multiple tweets of a single event based on pre-current and post event date . Results are given whether the event is positive or negative.

In future, we can evaluate the twitter user profile for each user(i.e. account creation date, number of tweets on the event) as another parameter to increase the scope of effect of event at

7. REFERENCES

[1] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin," A Practical Guide to Support Vector Classification", http://www.csie ntu edu tw ,web, July 22 2014.

[2] N Cristianini, J Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kemel-based Leaming Methods ", Cambridge University Press, 2000.

[3] Hiroshi Shimodaira,"Text classifying using Naive Bayes ", Document models, *http://www inf ed ac uk/teaching/courses/inf2b/learnotes/inf2b-learn-note07-2up*. 11 Feb 2014,web, 15 August 2014.

[4]Lindsay I Smith,"Principle Component Analysis", http://www.es otago acnz/cosc453/student tutorials/principal components, 2002, web. August 5 2014.

[5] Laura Auria, Rouslan A Moro, "Support Vector Machines ", 2008,web,20 August 2014.

[6] H Kim, P Howland, and H Park "Dimension reduction in text classification with support vector machines", Joumal of Machine Leaming Research,2005.

[7]Pang- Ning Tan Michael Steinbach, Vipin Kumar,"Introduction to Data Mining", pearson publications.

[8] Dan Jurafsky "Text Classification and Naive Bayes", The Task Of Text Classification, https://web stanford edu/class/cs124/lec/naivebayes .web. July 28 2014.

[9] http://help.sentiment140.com/for-students