International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-5, Issue-1)

Research Article | **January 2016**

Sentiment Analysis and Opinion Mining With Social Networking for Predicting Box Office Collection of Movie

Snehal. A. Mulay, Shrijeet J Joshi, Mohit R Shaha, Hrishikesh V Vibhute, Mahesh P Panaskar
Department of Computer Engineering and Information Technology, PVG's COET
Savitribai Phule Pune, University, Maharashtra, India

Abstract—

In recent years, social media has become pervasive and important for social networkingand content sharingBut then, the substance that is created from these sites remains to a great extent undiscovered. In this paper, we exhibit how online networking substance can be utilized to anticipate genuine results. System we propose will make use of Twitter Auth for streaming tweets. Tweets will be in unstructured format so our system will bring unstructured data into structured format for sentiment analysis. Once data brought into structured format, system will apply weights to tweets depending upon various criteria such as followers, following of actor, actress, director, producer and also rate of tweets on film hashtag. We further analyze sentiments extracted from Twitter which can be further utilized for forecasting movie box office collection of first week and overall collection

Keywords—Sentiment Analysis, Naïve Bayes Classifier, Hadoop, Map reduce, NLP, Big Data

I. INTRODUCTION

In this era social media is becoming more popular where netizens can express themselves, gives reviews etc. data generated through social media is nearly 10TB per day. With increase in such large amount of data it is necessary to develop a system which will make use of such large amount of data to perform analysis and predict future with social networking. So we are developing a system which makes use of twitter data for predicting box office collection of movie. This system include Natural Language Processing domain of computer science and scientific study of human language i.e. linguistics which is related with the interaction or interface between thehuman (natural) language and computer. Opinion mining or Sentiment analysis refers to a broad area of Natural Language Processing and text mining. It is concern not with the topic a document is about but with opinion it expresses that is the aim is to determine the attitude (feeling, emotion and subjectivities) of a speaker or writer with respect to some topic to determine opinion polarity. Initially it was applied for classifying a movie as good or bad based on positive or negative opinion. Later it expanded to star rating predictions, prediction of box office collection of movie.

A. What Is Sentiment Analysis?

Normal Language Processing is a space of software engineering and logical investigation of human dialect i.e. phonetics which is connected with the collaboration or interface between the human (natural) dialect and PC. Essentially NLP started as a sub-field of artificial intelligence. Supposition mining or Sentiment examination alludes to an expansive zone of Natural Language Processing and text mining. Sentiment analysis does not concerned with what the subject is about, but it concerns about what author wants to communicates . The point is to decide the state of mind (feeling, emotion and subjectivities) of a speaker or author as for some theme to decide supposition extremity. Initially it was applied for classifying a movie as good or bad based on positive or negative opinion. Later it expanded to star rating predictions, product reviews travel advice and other decision making processes. Sentiment Analysis recognizes the expressions in a content that bears some sentiment. The creator might talk about some target truths or subjective assessments. It is important to differentiate subjective and objective opinions. Conclusion investigation finds the subject towards whom the feeling is coordinated. A content might contain numerous substances yet it is important to discover the element towards which the slant is coordinated. It distinguishes the extremity and level of the assumption. Suppositions are named objective (realities), positive (signifies a condition of joy, ecstasy or fulfillment on part of the author) or negative (means a condition of distress, downfall or disillusionment on part of the essayist). The suppositions can further be given a score in view of their level of inspiration, antagonism or objectivity.

B. Challenges For Sentiment Analysis

*1) Implicit Sentiment and Sarcasm:*A sentence can be an implicit sentiment i.e. sentence may have opinion associated with it without having opinion deciding keywords.

Consider the following examples.

- I bought shirt a week ago, and it became fade after one wash.

The above sentence does not explicitly carry any negative sentiment bearing words although above statement is negative sentence. Thus identifying semantics is more important in Sentiment analysis than syntax detection.

*2) Domain Dependency:*Same word can give different polarity of statement which varies from domain to domain.

Consider the following examples.

- The story was unpredictable.
- The steering of the car is unpredictable.

In the first example, statement bears positive sentiment whereas the sentiment conveyed in the second is negative though both have 'unpredictable' word.

*3) Thwarted Expectations:*Sometimes the author deliberately sets up context only to refute it at the end.

Consider the following example:

- I decided to watch movie after I heard claims that it is best film, the actors and the supporting cast is good as well. However, it can't hold up.

In spite of the presence of more positive words that are positive in orientation the overall sentiment becomes negative because of the crucial last sentence.

*4) World Knowledge:*Often world knowledge needs to be incorporated in the system for detecting sentiments.

Consider the following examples:

- He is a Frankenstein.
- Just finished Doctor Zhivago for the first time and all I can say is Russia sucks.

The first sentence depicts a negative sentiment because Frankenstein is negative word whereas the second one depicts a positive sentiment. But to get correct sentiment one need to know about Frankenstein and Doctor Zhivago to find out the sentiment.

5) Subjectivity Detection: This is to differentiate between opinionated and non-opinionated text. Subjectivity Detection is used to enhance the performance of the system by including a subjectivity detection module to filter out objective facts. But this is often difficult to do. Consider the following examples:

- I hate drama movies.
- I do not like the movie "I hate stories".

The first example presents an objective fact whereas the second example depicts the opinion about a particular movie.

6) Entity Identification: A text or sentence may have multiple entities. It is extremely important to find out the entity towards which the opinion is directed.

Consider the following examples.

- Ios is better than android.

The examples are positive for Ios but negative for android and.

C. Application Of Sentiment Analysis

Mouth publicity is the process of conveying information from person to person and plays a major role in customer buying decisions. In commercial situations, consumers share attitudes, opinions, orreactions about businesses, products, or services with other people. People trust on families, friends, and others in their social network. Research also indicates that people appear to trust opinions from people outside their immediate social network, such as online reviews. This is where Sentiment Analysis comes into play. Availability of opinion rich resources like online review sites, blogs, social networking sites have made this "decision-making process" easier for us because we can get more reviews about product or services from consumers all across the world. With explosion of social networking platforms consumers have a power by which they can share opinions. Major companies have realized these consumer voices affect shaping voices of other consumers. Sentiment Analysis thus finds its use in Consumer Market for Product reviews, Marketing for knowing consumer attitudes and trends, Social Media for finding general opinion about recent hot topics in town, Movie to find whether a recently released movie is a hit.

II. EXISTING SYSTEMS FOR MOVIE PREDICTION

IMDB (Internet Movie Database)

IMDB is web based application which maintains all the information related to movies. This system also gives ratings to movie depending on reviews given by user. For giving rating to movie this system just finds the average of ratings given by user. In this system users create account and can rate movie from 1 to 5 stars. IMDB system then analyses these ratings to conclude which movie is hit, flop or average.

Advantage: This system recommends movie based on user reviews

Disadvantage: This system makes analysis over reviews submitted by user on their site only and doesn't make use of social network. As we have seen our system will overcome drawbacks of this system by making use of social media platform. Our system will also perform NLP and sentiment analysis to predict the mood of user towards movie.

III. METHODOLOGY

A. Research using twitter

Our system makes use of tweets as reviews given by user for particular movie. We are using twitter data because it is easily available through twitter API. Our system makes use of social hash-tags for streaming of twitter data. before streaming twitter data we need to create twitter application. Procedure for creating twitter application is as follows:

1. Creation of twitter application on twitter developer website. Name, Description, developer website and call back URL have to be specified in order to create app.
2. Twitter generates following details after successful creation of an application

Permissions: Each application on twitter has different permissions depending on the purpose. The twitter specifies the access level to indicate the permission the application can use.
Keys and Access Tokens: For identifying developer and the app Twitter generates 4 keys namely consumer key (API Key), consumer secret (API Secret), access token, and access token secret. To acquire data from Twitter or post updates on Twitter, the 4 keys have to be specified.

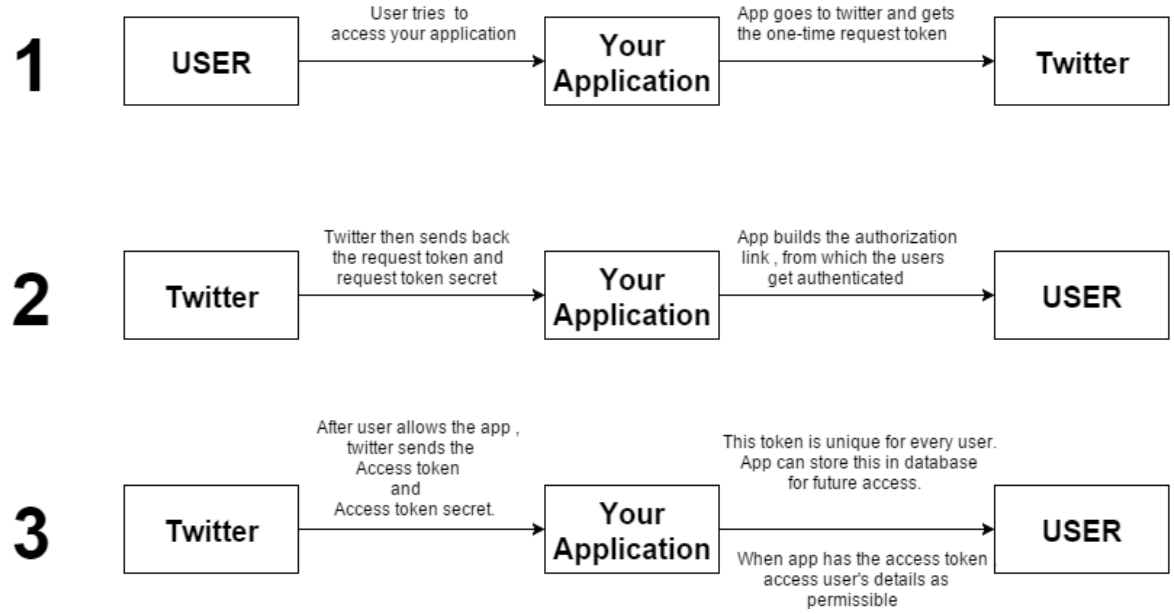

Fig.1 Tweet streaming using twitter API

B. Hadoop HDFS

Data available on twitter is of large amount in multiples of GB. If we try to process such a large amount of data with traditional database system then it will not be efficient. Traditional database system is suitable for structured data but it cannot handle unstructured data efficiently. Therefore we are using Hadoop HDFS file system along with mapreduce[1][2] technique for handling large amount of unstructured data.

C. Pre-Processing of Tweets

Natural Language Processing (NLP) is a technique that facilitates easy pre-processing of input text. Pre-processing refers to the cleaning and normalization of text to make sentiment analysis.

1. Words: Removal of stop words such as a, an, the, this which are very common and do not determine the sentiment of the text is carried out.
2. Punctuation: Punctuation marks such as commas and periods must be removed from the input text.
3. Duplicate Words: Duplicate words deviate the overall sentiment of the text. Duplicate words must be removed to make the input text for sentiment analysis.
4. Repeated Characters: Repeated characters in words such as “loooonnng” deviate the meaning of the original word. Thus words with repeated characters must be brought to their normal form.
5. Internet Acronyms and Emoticons: The use of emoticons and acronyms on the Internet such as ASAP, AFAIK prove a major problem while analyzing sentiment of the input text. A dictionary of common acronyms is maintained and cross-checked with the input text. The acronyms are expanded to their intended format.
6. URLs: Twitter API provides all the URLs present in the Tweet. The URLs do not change the sentiment of the input Tweet and thus must be removed from the Tweet. After treating the input Tweet with the above methods the input Tweet becomes ready for analysis. Only the required words which will make a difference to the sentiment are considered.

D. Naive Bayes Classifier

Naive Bayes classifier is a simple model for classification. It is simple and can also be used with text classification[3][9]. It is a probabilistic classifier based on Bayes' theorem with strong independence assumptions. This is the simplest form of Bayesian Network, in which all attributes are independent given the value of the class variable. This is called conditional independence. It assumes each feature is conditional independent to other features given the class. A Naive Bayes classifier is a technique that applies to a certain class of problems, namely those that phrased as associating an object with a discrete category. From numerical based approach group, Naive Bayes has several advantages such as simple, fast and high accuracy.

Bayes rule is given by

$\gamma(\alpha \mid \beta) = \gamma(\alpha)*\gamma(\beta \mid \alpha) / \gamma(\beta)$

Where α : Specific class

β : Document wants to classify

γ(α) and γ(β) : Prior probabilities

γ(α | β) and γ(β | α) : Posterior probabilities

Class α can be positive or negative class. Document is a review of particularmovie. The multinomial model of Naive Bayes captures word frequency informationin documents. The Maximum Likelihood Estimate (MLE) corresponds to the most likely value of each parameter given thetrainingdata. Equation for prior probability is given by following equation:

$\gamma(\alpha)= Nc / N$ …….(1)

Where Nc : The number of documents in class α

N : Total number of documents

Estimate the conditional probability γ(ω | α) as the relative frequency of term ω indocuments belonging to class α including multiple occurrences of a term in adocument.

$\gamma(\omega| \alpha) =(count(\omega,\alpha) +1) / (count(\alpha) | V |)$……(2)

Where

count(ω,α) : Number of occurrences of ω in training documents from class

count(α) : Number of words in that class

| V| : Number of terms in the vocabulary

Finally classify new document using posteriori probability. Let αNB is the posteriorprobability, αj is one of the class from class α and βi is ith document.

$\alpha NB = \arg\max_{\alpha j \in \alpha} \prod_{i} \gamma(\beta i \mid \alpha j)$……(3)

E. Sentiment Analysis

In Sentiment analysis first tweets will be streamed using hash-tag and stored in Hadoop database called HDFS. These tweets will be then pre-processed using NLP to get keyword which decides tweet as positive or negative.

F. Prediction

After sentiment analysis is performed, then weights are assigned to various factors such as casting, number of screens on which movie is releasing, number of tweets etc.which will contribute in movie success. The weights of all factors are combined together to predict overall success of movie over box office.

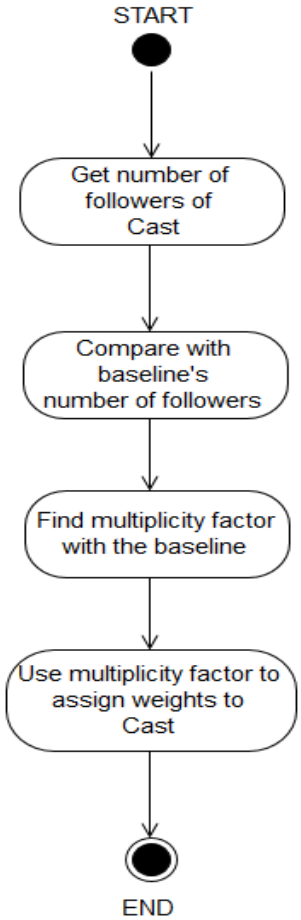

Fig.2 Predicing movie collection

IV. PROPOSED SYSTEM

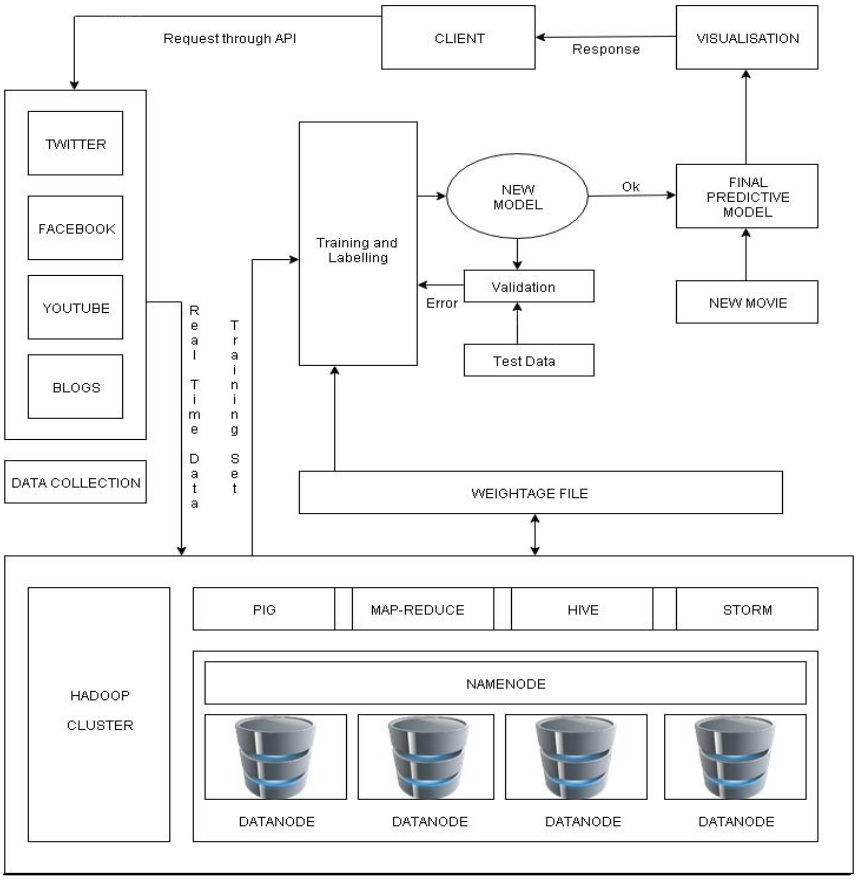

Fig.3 Architecture diagram of system

As shown in the above Architecture Diagram first user will stream tweets form twitter and store it in Hadoop HDFS file system. After data is stored in HDFS system will retrieve data with map-reduce technique. Then each tweet will be processed using NLP technique. After parsing tweets with NLP Naive-Bayes algorithm is applied to classify tweet as positive or negative. After this weightage file is created which consist of weights of all factors which affects movie success. Then weights of all factors are combined and applied to prediction model to predict box office collection of movie.

V. CONCLUSION

This paper discusses in detail the various approaches to Sentiment Analysis, mainly Machine Learning and Cognitive approaches. It provides a detailed view of the different applications and potential challenges of Sentiment Analysis that makes it a difficult task. Due to many challenging research problems and a wide variety of practical applications, it has been a very active research area in recent years. Feature engineering, as in several Machine Learning and Natural Language Processing applications, plays a vital role in SA. We have seen the use of phrases as well as words as features. It has been seen that Adjectives as word features can capture majority of the sentiment. Use of topic oriented features and Value Phrases play a significant role to detect sentiment when the domain of application is known.

ACKNOWLEDGEMENT

We take this opportunity to thank our project guide Prof. S.A. Mulay and Head of theDepartment Prof. G.V.Garje for their valuable guidance and for providing all the necessary facilities, which were indispensable in the completion of this project report. We are also thankful to all the staff members of the Department of Information Technology of PVG's College of Engineering and Technology, Pune for their valuable time, support, comments, suggestions and persuasion. We would also like to thank the institute for providing the required facilities, Internet access and important books.

REFERENCES

[1] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N, "Analysis of Big Data using Apache Hadoop and Mapreduce", in International Journal of Advanced Research in Computer Science and Software Engineering May 2014.

[2] Samira Daneshyar and Majid Razmjoo, "Large Scale Data Processing Using Mapreduce in Cloud Computing Environment", in international journal on web services computing December 2012.

[3] Lina L. Dhande and Dr. Prof. Girish K. Patnaik, "Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier", in International Journal of Emerging Trends and Technology in Computer Science August 2014.

[4] Jalaj S. Modha, Prof. Gayatri S. Pandi and Sandip J. Modha, "Automatic Sentiment Analysis for Unstructured Data", in International Journal of Advanced Research in Computer Science and Software Engineering December 2013.

[5] Vasu Jain, "Prediction of Movie Success using SentimentAnalysis of Tweets", in International Journal of Soft Computing and Software Engineering March 2013

[6] Sitaram Asur&Bernardo A. Huberman. (2010) Predicting the Future with Social Media. Proceedings the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, pp. 492-499

[7] L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie review mining and summarization," in Proc. 15th ACM Int. Conf. Inf. Knowl. Manage. 2006, pp. 43–50.

[8] Y. Lu, C. Zhai, and N. Sundaresan, "Rated aspect summarization of short comments," in Proc. 18th Int. Conf. World Wide Web, New York: ACM, 2009, pp. 131–140.

[9] Jayashri Khairnar and Mayura Kinikar,"Machine Learning Algorithms for Opinion Mining and Sentiment Classification", International Journal of Scientific and Research Publications (IJSRP), Volume 3, Issue 6, ISSN 2250-3153, June 2013.

