International Journal of Research in Engineering, Science and Management
Volume-2, Issue-4, April-2019
www.ijresm.com | ISSN (Online): 2581-5792

Box-Office Movie Prediction System

Apoorva Patil[1], Akshay Pujare[2], Monil Shah[3], Rohit Barve[4]
[1,2,3]*Student, Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, India*
[4]*Professor, Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, India*

***Abstract*: Predicting society's reaction to a new product in the sense of popularity and adaption rate has become an emerging field of data analysis. The multi-billion-dollar business, motion picture industry and there is a massive amount of data is available over the internet related to movies. This study proposes a decision support system for movie investment sector using techniques of machine learning. This research will help investors engaged with this business for avoiding investment risks. The system predicts an approximate success rate of a movie based on its profitability by analyzing historical data from different sources like International Movie Database, Rotten Tomatoes, Box Office Mojo and Metacritic. Using Support Vector Machine (SVM), Neural Network and Natural Language Processing the system predicts a movie box office will be a hit or flop based on some pre-released features and post-released features.**

***Keywords*: Movie prediction system**

1. Introduction

Data mining has become an important part in the world of Natural Language Processing, [1], [2] which deals with the interaction between computers and human languages and also deals with computers to program and easily process huge amount of natural languages. It is a difficult task for viewers to interpret the user generated content in the social media since it might mostly will be large in number. Nowadays, there is a drastic change in the social media sector from traditional media channels like magazines to social media channels like social networking sites it has lead researchers to study the scope of their exploitation in order to identify hidden knowledge contained within them. Millions of posts are appearing daily in web-sites that provide microblogging such as Twitter, Tumblr, Facebook. Authors of those posts share opinions on variety of topics and discuss current issues. As more and more users post about products they use, microblogging web-sites become valuable sources of people's sentiments. To build systems to mine opinions about any given topic, we need a powerful method for quickly identifying data that can be used for training. Sentimental analysis is the process of identifying positive and negative opinions, emotions, sentiments and evaluations. In these lexicons, entries are marked with their prior polarity: out of context, does the word seem to evoke something positive or something negative. Most of the present approaches to opinion mining and sentiment analysis are still far from being able to perfectly extract the affective information associated with natural language texts. Especially when dealing with social media, in which, contents are often very distinct and noisy, and the use of a domain-dependent training corpus is just not enough. The different types of sentiment analysis approaches are machine learning, lexicon-based, statistical and rule-based approaches. To determine the sentiment by training known dataset, the "Machine Learning method" uses several learning algorithms. It emphasizes on automatic methods. In other words, the goal is to devise learning algorithms that do the learning automatically without human intervention or assistance.

- The "lexicon-based approach" includes calculating sentiment polarity using the semantic position of words or sentences in the review.
- The "semantic orientation" is a part of subjectivity and opinion in text. It is a database of lexical units for a language along with their sentiment orientations. This can be indicated as a set of tuples of the pattern (lexical unit, sentiment). Here, the lexical entity may be words, word senses, phrases.
- The "rule-based approach" looks for opinion words in a text and then classifies it based on the number of positive and negative words. It considers different rules for classification such as booster words, emoticons, negation words, mixed opinions, dictionary polarity, idioms, etc.

A. Aim and objective

The large growth in number of movies releasing over the past few decades Movie Prediction is necessary. The only way people can check whether the movie will be worth to watch is through applications, so this system would analyze the reviews posted by other users, as these reviews are large in number which the user cannot read and gets confused. Following are the aims and objectives suggested by our system.

- To evaluate the individual opinion regarding movies.
- To evaluate the emotional tone behind the series of the word.
- To teach machine to analyze the various grammatical nuances.
- To implement an algorithm for automatic classification of text into positive or negative comments.

B. Problem statement

Since now –a –days there are many good movies and bad movies or we can say some of the movies are neutral, so many

a times people realize and analyze after seeing a movie that they have wasted their time and money. In today's scenario it is difficult to analyze whether the particular movie will be a good, bad or neutral movie before seeing it or before it gets released.

2. Literature survey

There are two methods broadly used to recognize the assessments from the content. They are Symbolic methods and Machine Learning procedures [3].

- *Analysis using Symbolic Techniques:* A typical system utilizes the accessibility of lexical assets. Turney recommended an approach for opinion investigation called 'sack of words'. In the specified approach, singular words are dismissed and just accumulations of words are considered. He accumulated word having descriptors or qualifier for the extremity of survey from a web crawler Altavista. A lexical database called WordNet was utilized by Kamps et. al. which decides a passionate matter in a word. WordNet conveys equivalent words and separation metric to discover the introduction of descriptors. To beat impediments in lexical substitution assignment, Baroni et. al. built up a framework upheld by word space show formalism along these lines speaking to neighborhood words. EmotiNet adroitly spoke to the content that put away the structure of genuine occasions in a space. This was presented by Balahur et. al.
- *Analysis using Machine Learning Techniques:* Under this system, there are two sets, to be specific a preparation set and a test set. By and large, the dataset which is gathered from various sources and whose conduct and yield esteems are known to us falls into the classification of preparing informational collections. Conversely with this, the datasets whose esteems or conduct are obscure to us are called as test informational indexes. Here various classifiers are prepared with preparing information and after that obscure information or we would say be able to a test information is given to this model to get coveted outcomes. Machine Learning comprises of different distinctive classifiers, for example, Ensemble classifier, k-means, Artificial Neural Network and so on. These are utilized to characterize surveys. Y. Mejova et al in his exploration work suggested that we would use be able to the nearness of each character, recurrence of events of each character, word which is considered as nullification and so forth as components for making an element vector. He additionally demonstrates that we can successfully utilize unigram and bigram ways to deal with make include vector in Sentiment investigation. Domingos et al proposed that Naive Bayes functions admirably for subordinate components for certain issue.

3. Proposed system

Predicting society's reaction to a movie in the sense of adaption rate and popularity has become an emerging field of data analysis. The system predicts a movie box office will be a hit or flop based on some pre-released features and post-released features, using Support Vector Machine (SVM), Neural Network and Natural Language Processing. We are going to perform sentiment analysis in our project to know people's opinion. Sentiment analysis is basically concerned with opinions from text and analysis of emotions. We can refer sentiment analysis as opinion mining. Sentiment analysis justifies and finds the person's sentiments with respect to a given source of content. Social media contains massive amount of sentimental data in form of blogs, tweets and posts, updates on the status, etc. Sentimental analysis is very useful to express the opinion of the mass of this largely generated data. Twitter sentiment analysis is tricky as compared to broad sentiment analysis because of the slang words and repeated characters and misspellings. We know that the maximum length in Twitter is 140 characters of each tweet. So it is important to recognize correct sentiment of each word. In our project we are proposing a model with high accuracy of sentiment analysis of tweets with respect to latest reviews of upcoming Bollywood and Hollywood movies. Most of the present approaches to opinion mining and sentiment analysis are still far from being able to perfectly extract the affective information associated with natural language texts. Especially when dealing with social media, in which, contents are often very distinct and noisy, and the use of a domain-dependent training corpus is just not enough.

4. Methodology

Various methods have been used to do sentiment analysis of tweets. In our research we have used two approaches. Lexicon Based and Rule Based approach. Matching algorithm is used to find out the polarity of words. We are classifying these tweets as positive and negative to give sentiment of each tweet. The following figure shows the entire proposed system architecture. The proposed system contains various phases of development. A dataset is created using twitter posts of movie reviews. Here we perform sentiment analysis on each word in the tweets. First we collect all the tweets in a file. Then in the next step, extract sentences based on punctuation mark. Then we tokenize sentences based on word boundaries. In the next step extracted words are matched against word in the positive and negative word corpuses that we have already built. Polarity of each word is thus found out using matching algorithms. Then in the next phase, final polarity of sentence is drawn based on the count. In the final phase the polarity of entire tweets are found using the count and generate a summary based on that.

A. Creation of dataset

- A dataset is created using twitter posts of movie reviews and related tweets about those movies.

- The below table shows dataset used for (tweets) used for testing.

Table 1
Statistics of Dataset

Dataset	Total
Training	1000
Testing	500

Tweets for testing are selected by eliminating all tweets containing words other than meaningful words in English. Thus, we got 500 tweets out of 1000.

B. Extraction and tokenization

From all the tweets collected for testing, each tweet is separated from the file by keeping punctuation mark as the base. After that each word is extracted from the tweet by considering space between the words as its base. Each extracted word is called as a "token".

C. Creation of Word Corpuses

A positive word corpus contains all possible positive words which are usually used in tweets similarly a negative word corpus is also created.

D. Matching Algorithm

A matching algorithm is used to match words or sentences in a file against words or sentences in the corpuses. Since the words are separated from sentences, it is easy to compare that with each word in the two corpuses. This is nothing but the process of doing Lexicon Based Approach for calculating the polarity. Thus we will be able to find out the polarity of each word in the sentence. This process is repeated for the entire sentence in document.

5. Data flow Diagrams

- DFD level 0

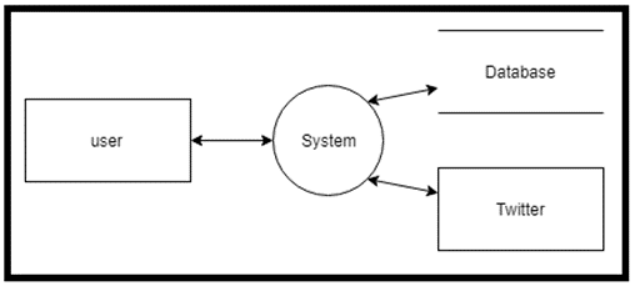

Fig. 1. DFD level0 for box office movie prediction system

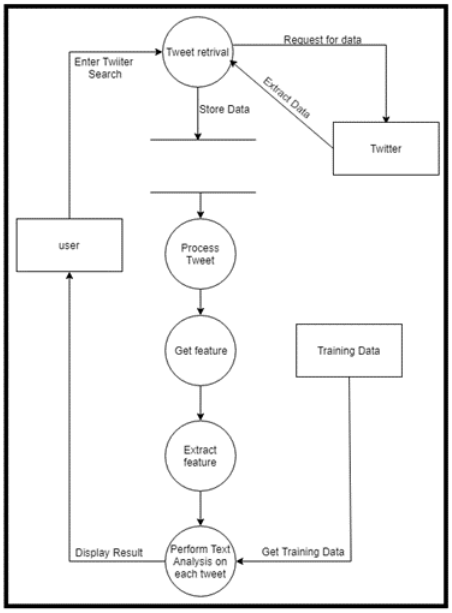

Fig. 2. DFD level 1

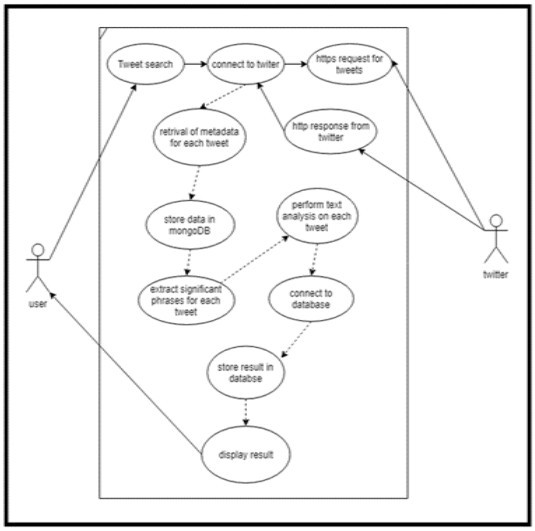

Fig. 3. Use case diagram

6. Conclusion

We did some preliminary study in using sentiment analysis to predict a movie's box office success. The results show that the box office success can be predicted by analyzing sentiment of the movies with simple metrics and pretty good accuracy. We understand that there might be more than one factor which affect the movie box office success, but we concentrate on sentiment analysis in this work. As sentiment analysis on twitter itself is a challenging topic, we feel that there is a long list of future work. However, this problem itself is an interesting and promising area.

References

[1] Akshay Amolik, Niketan Jivane, Mahavir Bhandari, M. Venkatesan, Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques, IJET (2014).
[2] Vidya D. Gonugade, Shanmukhappa A. Angadi, Sentiment Analysis on News Article through Symbolic Data Analysis, IJRTER (2016).
[3] Ms. Md. Sania Sultana, Opinion Mining on Twitter Data of Movie Reviews, IOSR journals (Jul.-Aug. 2017)
[4] Waterfall-model: www.tutorialspoint.com/sdlc/waterfallmodel
[5] Measuring Audience Sentiments about Movies using Twitter and Text Analytics, Analytics Vidya.
[6] Bijith MaraKarkandy, "Using twitter data to predict the performance of Bollywood movies (Sept. 2015).
[7] Chin-Chang Chand, Prediction of movies box office using social media, (August 2013).
[8] Willis R, "Cost analysis: Cost analysis and estimating–Tools and techniques."