Movie Exploration System

Aditi Kacheria
Student
K. J. Somaiya College of Engineering
Department of Computer Engineering

Kanak Jain
Student
K. J. Somaiya College of Engineering
Department of Computer Engineering

Nidhi Shivakumar
Student
K. J. Somaiya College of Engineering
Department of Computer Engineering

Shreya Sawkar
Student
K. J. Somaiya College of Engineering
Department of Computer Engineering

ABSTRACT

Movies nowadays are not restricted only to the elite but are also for the masses. With multiple movies releasing every week, it is difficult for people to choose the movies that are worth watching. Sometimes people want a very general opinion on the movie, rather than a critic's review. Apart from this, people may also be interested in watching movies of a particular genre based on their liking.

To obtain a public opinion, data from various sources, like Twitter, Facebook can be extracted and thoroughly studied to infer the rating of a movie. To obtain suitable recommendations, the user's activity can be monitored to provide the user with appropriate movies. Movie Exploration System (MES) combines all the above mentioned ideas to give efficient results to the users.

General Terms

Movie Exploration System, Ratings, Recommendations

Keywords

Sentiment Analysis, Twitter, Naïve Bayesian.

1. INTRODUCTION

Many existing movie rating systems like IMDB, Netflix provide a rating and an overall review of a movie based on the critic's opinion. However, many critically acclaimed movies do not necessarily do well at the box office and do not have a huge liking amongst the general public. At the same time there are movies which are talked about and liked by the public but they do not go down well with the critics. In recent years, social media networking has provided a platform for people to express their opinions on current issues like politics, latest movie releases etc. These opinions, in the form of tweets, facebook posts, likes etc. can be used to analyse the public's view of any subject. This approach which uses the public's opinion and not the critics view can be used to gauge the rating of a movie.

In fact, the movie business is no longer restricted only to entertainment. A lot of businesses are closely associated with this industry. These investors look out for movies which promise a high box office revenue. Hence they cannot go with a critic's rating of the movie which may be personalized and influenced.

Therefore developing a Movie Exploration System (MES) which incorporates the following features will fulfill the need of the hour.

1) Ratings for latest releases using some form of public data source, like tweets.
2) Personalised movie recommendations.

Users of the system can be divided into different classes -

1) General users that want recommendations and ratings.
   a) Guest user who will be recommended top ‘n’ movies based on the genre provided by him.
   b) Registered user (new/ inactive) who will provide the system with certain static details, such as name, genre. MES will recommend movies to the user based on the static profile.
   c) Active user whose movie ratings will be monitored in order to capture any changes in the genre watched by the user, so that the recommendations can be made based on the dynamic profile of the user.
2) Producers or investors that want to check the rating of the movie.

The paper is organized as follows. The next section gives the related work. The third Section describes the approach used for MES. The fourth section highlights the results and conclusions derived from the experiment and the last section provides the future enhancements that can be applied to MES.

2. RELATED WORK

The rise of social media in recent years has led to its extensive use in the field of natural language processing and machine learning. As a result, sentiment analysis is one such area which is being explored tremendously. [1] gives a detailed summary of Twitter Sentiment Analysis Methods and limitations of tweets as a dataset for analysis. It also provides us with twitter specific features like hashtags, emoticons, @ etc. However, [2] explains how the use of emoticons affects the accuracy of the system. [3]When it comes to rating a tweet that talks about a movie, we have to keep in mind the critical period, i.e. mostly 2 weeks before and after the release date. This is when the movie is most talked about. This points out a very important feature that can be used to classify tweets, i.e. the date on which the tweet was written. [4][5] list a number of ways to collect data sets from twitter and label them. The papers give a description of the accuracy attained using various combinations of features and classifiers. Common classifiers used for this purpose are Naive Bayes Classifier, Support Vector Machine [3]. [6] Hadoop is an efficient and scalable open source framework that processes big data in a distributed manner. This uses naive bayes to carry out the study of classification.

3. METHODOLOGY

3.1 Rating Generator

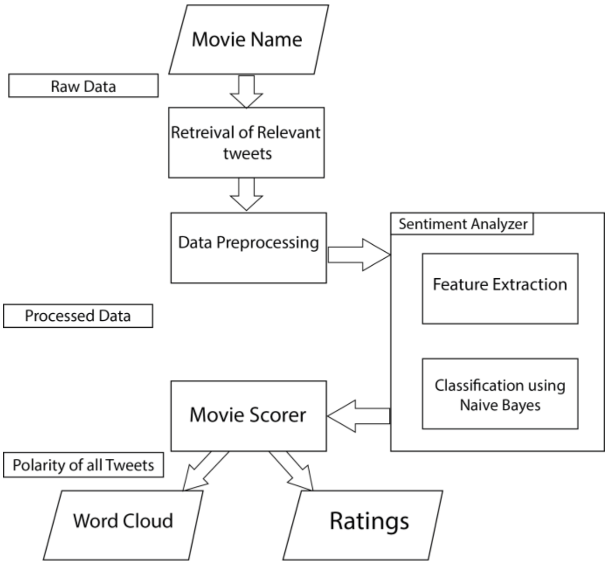

Fig. 1 Rating Generator System Architecture

MES focuses on sentiment analysis to provide ratings. Sentiment analysis is a very popular approach that uses Natural Language Processing (NLP) and Machine Learning (ML) techniques for recognizing the sentiment of textual data. Following steps are involved in the generation of ratings.

1) **Data retrieval** - Using the searchTwitter (movie_name) command in R, we retrieve all the tweets relevant to the movie.

2) **Data preprocessing** - Tweets are first converted into lowercase. However, the raw corpus from R contains redundant data that needs to be filtered. This involves removal of:

   - Hashtags
   - Emoticons
   - Punctuations (except '!')
   - Http links

   This stage also replaces of words like 'gud' with 'good', i.e. finding the root word from slang word [7].

3) **Feature extraction** - Following are features MES uses for classification [8]

   - Unigram
   - Bigram
   - Date
   - POS tagger
   - Favorited: This is a twitter specific feature that we have taken into consideration. If the tweet is favorited it is the opinion of not just one but more people. Hence we take it as a feature.
   - Emphasis: This is a social media related feature that we have included. The sentiment of a word with emphasis is different from a word without emphasis.

4) **Calculation of various features** - We show the calculations using an example tweet. We take the following tweet: #Madaari must watch amazing movie this weekend!!! congratulation to the entire team!!! @irrfank #NishikantKamat

   The unigram value is calculated as:
   Watch – 0, Weekend – 0, Amazing – 1, Congratulation – +1, Entire – 0, Team – 0
   Total = 2/6 = 0.33
   The bigram value is calculated as:
   Watch amazing- 1, Amazing weekend – 1, Weekend congratulation – 1, Congratulation entire – 1, Entire team – 0
   Total = 4/5 = 0.8
   The POS value is calculated as:
   Amazing – 1
   Total = 1
   The date value is calculated as:
   If (current_date and date_of_tweet are 2 weeks apart)
   Date = 5
   Else
   If (current_date and date_of_tweet are one month apart)
   Date = 3
   else Date = 1
   The emphasis is calculated as:
   Weekend!!! – 1, Team!!! – 1
   Total = 2/5

   This shows the calculation of the features. The next part shows how to classify a tweet after we get the tuple consisting of the above features.

5) **Classification -** We have used the Naïve Bayesian Classifier [3]. The tweet is represented by the features listed above. The classifier classifies tweets into two categories, positive and negative. P(X|Yes,No) = P(A1|Yes,No) * P(A2|Yes,No)......*P(An|Yes,No)

   Where A1, A2,...An are the various attribute and Ai >/=/< c where c is a constant and "Yes" means that the tweet is positive and "No" means that its negative.

   Depending on the values of P (X|Yes) and P (X|No), whichever is greater, the tweet will be classified as positive or negative.

6) **Rating Generator** - Rating = (Number of positive tweets) / (Total number of tweets) * 10

3.2 Movie Recommendation

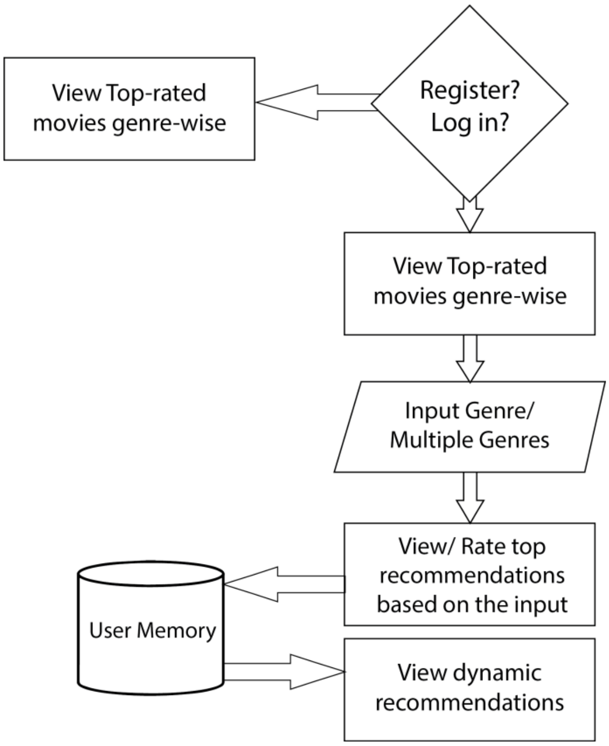
View Top-rated movies genre-wise
Register? Log in?
View Top-rated movies genre-wise
Input Genre/ Multiple Genres
View/ Rate top recommendations based on the input
User Memory
View dynamic recommendations

Fig. 2 Recommendation Generator System Diagram

MES also deals with recommending movies to the users based on their interest. For creating the approach for recommendation we have used the MovieLens dataset [9]. This consists of the following tables:

1) movies - movieid, title, genre.
2) ratings - userid, movieid, rating.

During registration, we ask the user to provide us with certain information such as his/her name, age, email ID, genre (that the user is interested in). Based on this, we recommend the user the top movies that align with his interest. Initially, the movies are all divided into clusters based on the genre that the movie belongs to. The information in the ratings table is present in the following format: userid, movieid, ratings and genres. Each movie can belong to multiple genres based on which clusters are formed. For eg., Deadpool belongs to the following genres: Action and Comedy. Each cluster has a unique cluster ID. Similarly, each user also has a unique user ID. Initially, the user ID is linked with the cluster ID of the genre that interests the user. As and when the user starts rating movies, the system keeps track of the genre of these movies, thereby tracking the user's interest at all times. If there is a change in the genre of movies that the user is rating, the user profile is updated with the cluster ID of the new genre. A user can have multiple cluster IDs linked to his/her user ID at any given point of time. If at all the user is linked to cluster of genre 'A' but starts deviating towards another cluster of genre 'B' then, over a period of time, if the user completely loses his interest towards genre 'A', his profile will also get updated and his user ID will no longer be linked with the cluster ID of genre 'A'.

4. RESULTS AND CONCLUSION

The accuracy of the ratings generated by the system was obtained by comparing the system generated label given for the tweet with the manually labeled tweets. After comparing these two labels, our accuracy was around 60%. Out of the 200 manually labeled tweets, 118 were labeled correctly by the system. The recommendation engine can be tested by comparing the recommendations generated by our system and that of established systems like Netflix.

5. FUTURE ENHANCEMENTS

1) System must be able to capture the sarcasm in tweets.
2) System must be able to suggest a suitable lead pair based on the scripts' demands, budget of the movie, etc.
3) Addition of a trending column where current movies can be shown to the user by extracting from various data sources.
4) System can be made more localised if regional movies are also included in the dataset.

6. REFERENCES

[1] Peter D. Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424.

[2] Zhi-Dan Zhao and Ming-Sheng Shang, User-based Collaborative-Filtering Recommendation Algorithms on Hadoop, 2010 Third International Conference on Knowledge Discovery and Data Mining.

[3] Alexander Pak, Patrick Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Universit ́e de Paris-Sud, Laboratoire LIMSI-CNRS, Bˆatiment 508, F-91405 Orsay Cedex, France

[4] Vasu Jain, Prediction of Movie Success using Sentiment Analysis of Tweets, The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13]-Vol. 3, pp. 308-312

[5] Vishal A. Kharde, S.S. Sonawane, Sentiment Analysis of Twitter Data: A Survey of Techniques, International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016, pp. 6-15

[6] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford.

[7] Akshay Amolik, Niketan Jivane, Mahavir Bhandari, Dr.M.Venkatesan, Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques, International Journal of Engineering and Technology (IJET), Vol 7-No. 6, Dec 2015-Jan 2016, pp 2038-2044.

[8] Anastasia Giachanou and Fabio Crestani, Like It or Not: A Survey of Twitter Sentiment Analysis Methods, ACM Comput. Surv. 49, 2, Article 28 (June 2016), 41 pages.

[9] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages.