Trends and Sentiment Analysis of Movies Dataset Using Supervised Learning

Shweta Taneja, Siddharth Bhasin, and Sambhav Kapoor

1 Introduction

The motion pictures such as movies and tv shows can be classified into categories which are called genres. However, a movie may have more than one genre. Grouping the movies into broad categories of the genre is yet another challenging classification task to be performed. This activity of grading movies in classes helps both the viewers as well as the critics to draw various conclusions. Labeling the movies into different genres gives more clarity on the type of viewers it shall attract. These trends and results on the genre preferred more by the people will also help the film directors and creators in making the films or shows. In this work, a method of movie classification and sentiment analysis has been proposed.

In our research, the concept of data classification algorithm is used [1]. Classification, being a supervised learning method in machine learning, helps to map the observation to a set of categories which, in this task, helps to identify the genre of a particular movie with the help of its description. Some famous applications of classification are the spam-ham classification of a given email, diagnosing a patient on the basis of the observed characteristics (certain symptoms, blood pressure value, sex etc.) in hospitals etc.

In our work, classification helps in grouping the descriptions of movies into six broad categories or genres which are drama, horror, comedy, western, thriller

S. Taneja (✉) · S. Bhasin
Department of Computer Science, Bhagwan Parshuram Institute of Technology, Guru Gobind Singh Indraprastha University, Dwarka, Delhi, India
e-mail: shwetataneja@bpitindia.com

S. Kapoor
Department of Instrumentation and Control Engineering, Netaji Subhas University of Technology, Dwarka, Delhi, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

B. Agarwal et al. (eds.), *Proceedings of International Conference on Intelligent Cyber-Physical Systems*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-16-7136-4_25

and documentary. The endeavor is to create a system that can perform rigorous classification.

Multilabel classification techniques involve assigning an instance of an attribute to more than one class label. News articles are a common example of this.

Following are the classifiers that are mostly used in multilabel classification.

One-versus-rest (OvR)

One-versus-rest (OvR) is also called as One-versus-all (OvA) method [2]. In this, a real-valued confidence score is created by the base classifier for its decision, instead of a class label. OvA learner constructed from binary classifier performs a training algorithm where inputs are a learner L, samples X, labels y where yi $\in$ {1, …, K} (for some sample Xi) and the output is a list of classifiers f(k) for k $\in$ {1, 2, …, K}. In order to predict the label k for a classifier, we apply the classifiers to an untold data sample x that gives the maximum confidence score:

$$\hat{y} = \text{argmax} f_k(x), k \in \{1 \ldots K\} \tag{1}$$

OvR with Support Vector Machines

SVM alone supports only binary classification. Therefore, in order to handle the separation of multiple classes, essential parameters and constraints are also added in these extensions [3].

OvR with Logistic Regression

The logistic function is an S-shaped curve (like the sigmoid function) which helps in mapping the values between 0 and 1 by taking real numerical values. It uses Euler's number (e) which is the base of the natural logarithms [4]. Logistic regression is a linear method that predicts the probability and transforming it using a logistic function. The equation can be represented as

$$1/(1 + e^{-\text{value}}) \tag{2}$$

Binary Relevance with Gaussian NB

Binary relevance (BR) is considered as the main baseline for classification in machine learning. The BR method is based on the assumption of independent labels. Hence, the classifier studies each label independently and declares it as irrelevant or relevant. According to several matrices, BR is not only effective in producing ML classifiers but is also computationally efficient. BR together applied with Gaussian Naive Bayes stimulates the model for the multilabel classification. Naive Bayes is based on the principle of MAP (maximum a posteriori) [5]. It is an efficient and popular classifier.

Label Powerset with Logistic Regression

It is used in multilabel classification [6].

Sentiment analysis is an area under natural language processing (NLP) that helps in identifying the sentiment within a text. It is, therefore, also known as opinion mining [7]. Sentiment analysis works on the unstructured data of raw texts and converts them into structured data that can be useful for any brand, organization, politics etc. In our work, we have applied sentiment analysis on the tweets of the Twitter users to find out which genre of movies/tv shows the users like the most and classified the tweets as positive, negative or neutral with the help of TextBlob. TextBlob is a powerful python library that can be used for sentiment analysis and offers a simple API to access its method and accomplish standard NLP operations. TextBlob analyzes an English phrase in the form of a score. Each lexicon has the scores for polarity, subjectivity and intensity with their different specified ranges. The polarity defines if the sentiment for the text is positive, negative or neutral which helps us to understand what people actually think related to movies of a particular genre. In this way, while implementing sentiment analysis we get a general public view over the Twitter platform of the most favorable genre or the trending genre and movies of those genres that can have praising outcomes on the screen which can help in a good critic rating or even can do a good business.

TextBlob is a powerful NLP library for python that can be used for sentiment analysis. It helps in determining the polarity and subjectivity of the text. We have applied TextBlob to the text of tweets to find out about their polarities [8]. The polarity using the TextBlob ranges from -1 to 1. We classified all the tweets whose polarity < 0 as -1 and the tweets whose polarity > 0 as $+1$. Then we calculated the number of positive, negative and neutral tweets in the dataset. After that, we computed the percentage of positive and negative tweets using the formula:

$$\text{Percentage of positive tweets} = \left(\frac{\text{Number of positive tweets}}{\text{Total number of tweets}}\right) * 100 \quad (3)$$

Similarly, for negative tweets,

$$\text{Percentage of negative tweets} = \left(\frac{\text{Number of negative tweets}}{\text{Total number of tweets}}\right) * 100 \quad (4)$$

This research work focuses on analyzing trends and sentiments of different movie genres. It covers on following points:

- Using hybrid algorithms to differentiate between multiple labels of the movies. This classification is achieved by using multiple hybrid algorithms such as pipeline for applying SVM with one-versus-rest classifier, binary relevance and Gaussian NB and classifier chains with logistic regression.

- These algorithms are compared by measuring the accuracy.
- For each subsequent genre tweets are extracted using the Twitter API.
- Twitter data is mined for making interpretations.

Sentiment analysis of the genres is done from the extracted tweets. This helps to get insights like evaluating the viewpoint, evaluations, and feelings of a speaker/writer on the social media platform.

The paper is organized as follows: Sect. 2 gives the state of the art in this field. Section 3 highlights the proposed work, which is divided into two subsections. The proposed work includes the flowchart along with pseudo-code. Section 4 shows the dataset used in the work. Section 5 shows the results followed by the conclusion.

2 Related Works

Table 1 shows the summary of work done by different authors.

3 Proposed Work

The proposed methodology is divided into two subsections, namely classification of movie genres and sentiment analysis of the genres.

3.1 Classification of Movie Genres

The flowchart in Fig. 1 shows the stepwise procedure performed in the classification of movie genres.

In the first step, the data extracted from Kaggle is saved into a csv file for further processing. We have considered three attributes of the data stored, that is, the movie name, description and movie genre, out of which description and genre are used in the data classification system. The second step is the most crucial one which is the data pre-processing. In order to influence the results and analysis, pre-processing and mining the data is one of the most crucial parts. The first part in data pre-processing is creating dummies. Here we build a dummy variable or column for each categorical value in the genre. In this way, we store the numerical value against each description representing the genre. Cleaning the data is the process in which initially, the text is tokenized that is segmented into clauses or words, we clean text by removing the unnecessary data in it which may include tags, punctuations, links, emails, phone numbers and other multiple pointless words which are not essential in categorizing the data and does not help the model. Stop words are the words present in the NLTK corpus which are the commonly used words and are unlikely to be useful for

Table 1 Summary of contributions of different authors

S. no.	Authors	Contribution
1	Kadam et al. [9]	The authors have used sentiment analysis as a powerful tool to find the most common features in a program that users generally like in order to increase the success rates of newly proposed TV programs
2	Mhaigaswali and Giri [10]	The authors have stated the importance of social networking in predictive and descriptive analytics
3	Rahim et al. [11]	The authors have stated the importance of YouTube for videos and have used the movie trailers data
4	Zubiaga et al. [12]	The paper has focused on identifying the trends on Twitter by looking at the earliest tweets that produce the trends and categorizing the trends early on. It uses language-independent features relying on the social spreads of the trends to segregate among the trending topics
5	Satyavani et al. [13]	K-Means algorithm has been used by the authors to compare different TV shows on the basis of their popularity
6	Schmit and Wubben [14]	Different classification and regression techniques are used by the authors on tweets from Twitter to make a prediction about the rating of newly released movies. Also, this paper focuses on the importance of textual features in the predictive machine learning tasks
7	Wang and Zhang [15]	The authors have used the Gaussian kernel support vector machine (SVM)model to predict the movie genre preference using customers' behavioral, demographic and social information. Different VC (VapnikChervonenkis) dimensions are used in this paper to compare the error-out-samples
8	Battu et al. [16]	The authors have created a Multi-language Movie Review Dataset and used different techniques such as SVM, random forest, FCNN, LSTM, GRU, hybrid models to predict the movie genre
9	Maloof [17]	The author has shown the application of machine learning in digital investigation
10	Khan and Urolagin [18]	The paper depicts the importance of machine learning in social media data
11	Bhardwaj et al. [19]	For sentiment analysis, a novel approach has been developed for the travel industry
12	Jindal and Taneja [20]	In this paper, the authors have developed an algorithm for multilabel categorization of text documents. It is implemented on five text datasets and has shown promising results

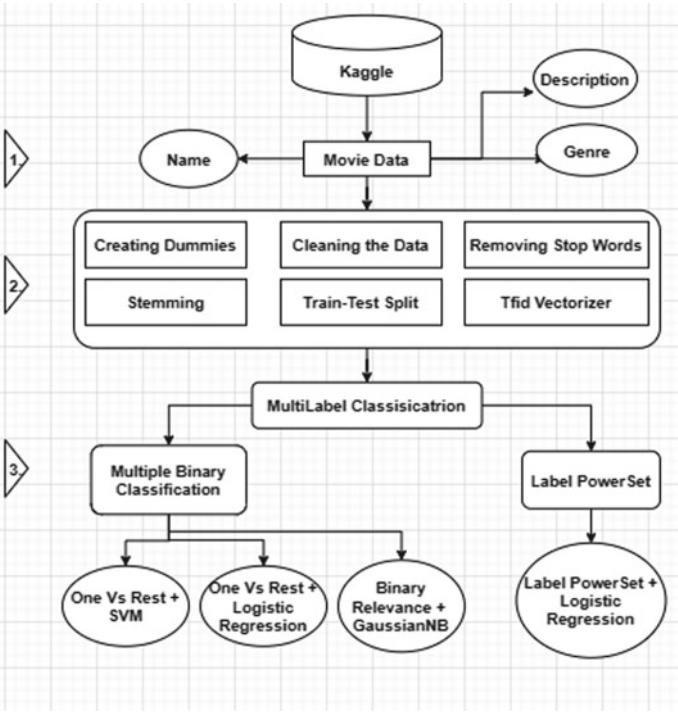

Fig. 1 Flowchart of classification of movie genres

learning. Stemming is the process of generating morphological variants of a base word. A stemming algorithm is the one that reduces or replaces the words to their root word or a common stem. For instance, likes, liking, likely or liking are reduced to like which is their root word.

Stemming helps to reduce redundancy. The data is then segregated into sets: training and testing. Next is the TfidVectorizer, Tfid is the term used for term frequency–inverse document. In TF-IDF, term frequency replicates how often a word appears within a document and marks its frequency and inverse document frequency downscales or removes words that appear a lot across the text. After pre-processing the data, the third step comes in which the system is built for multilabel classification. Since against each movie description we have multiple genres attached to it, we have implemented hybrid algorithms in order to achieve the task of classification. Hybrid algorithms are basically a way of combining models and bringing together the strengths of both knowledge representations.

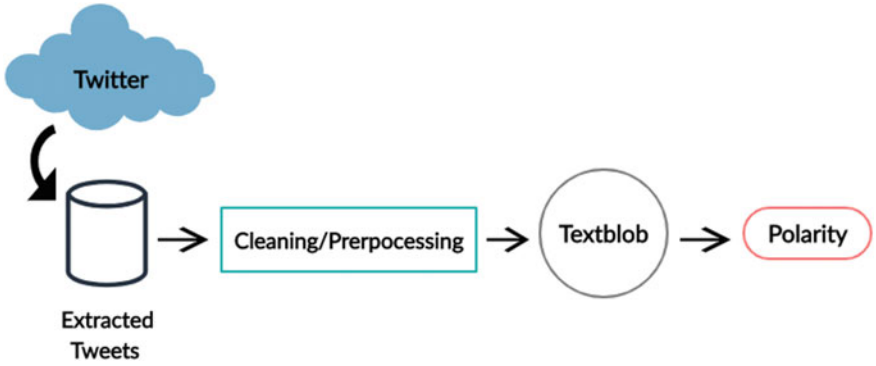

Fig. 2 Flowchart for sentiment analysis starting from extracting Twitter data to the polarity of the genre

3.2 Sentiment Analysis of Movie Genres

The second subsection is sentiment analysis of movie genres. The algorithm used for the sentiment analysis of the tweets using TextBlob is given as follows:

1. Use Tweepy to extract tweets from Twitter.
2. Using hashtags of different genres, tweets are extracted for those genres.
3. Merge all the tweets extracted to prepare a dataset consisting of the username along with the text of the tweet.
4. Pre-processing is done to clean all the non-letters in tweets.
5. Analyze the tweets for sentiment analysis using TextBlob.
6. The polarity of each tweet is then found out as $+1$, -1 or 0 and is added to the new dataset.
7. The percentage of positive and negative reviews for each genre is then calculated (Fig. 2).

The pseudo-code for the sentiment analysis is given below. The code demonstrates the method to find the polarity of the sentiment for a given text of lines. Once the polarity of the text is determined by TextBlob, the outcome of positive and negative reviews in tweets is shown and printed for the results.

```
l = list()
for t in texts:
    analysis = TextBlob(t)
    if analysis.sentiment.polarity> 0:
l.append(1)
elifanalysis.sentiment.polarity == 0:
l.append(0)
    else:
l.append(-1)
act = pd.DataFrame({"tcexts":texts,"polarity":l})
sum1=0
for polar in act["polarity"]:
    if polar==1:
        sum1=sum1+polar

sum2=0
for polar in act["polarity"]:
    if polar==-1:
        sum2=sum2-polar
print("The    percentage    of    positive    reviews    is
",(sum1/len(act))*100)
print("The    percentage    of    negative    reviews    is
",(sum2/len(act))*100)
```

4 Dataset Used

For the implementation part, we have used the Movie Dataset from Kaggle [21]. It contains metadata of 45,000 movies present in the Full Movie Lens Dataset. This dataset contains 26 million ratings from users. These are rated are on a scale of 1–5 (Table 2).

Table 2 Dataset used

S. no.	Description	DislikeCount	LikeCount	Title	VideoId	Genre
1	When sibblingsjudy and	84	36,526	Alexa & Katie	AN-Wmg8ByVI	Comedy
2	A family wedding…	94	1356	Tidelands	vI03_g-hlGM	Crime
3	Never lose sight of	7300	156,986	Bird Box	o2AsIXSh2xo	Thriller
4	What happens when …	1609	93,367	The Princess Switch	sP_-iNVjiSE	Romcom

5 Results

Python language has been used for the implementation of the work. Figure 3 shows the percentage of positive reviews tweeted by Twitter users for each genre. According to the graph, the movies/shows of comedy genre received the highest percentage of positive reviews followed by western, thriller, action, documentary, horror and drama.

Figure 4 shows the percentage of negative reviews tweeted by Twitter users for each genre. According to the graph, the movies/shows of the horror genre received the highest percentage of negative reviews followed by thriller, drama, western, action, comedy and documentary. As per the experiment results, the comedy genre is seen to have the highest percentage of positive tweets (66.04%), then western (56.4%), thriller (48.3%), action (47.92%), documentary (40.15%), horror (37%) and drama (23%). The horror genre was observed to have the highest percentage of negative reviews (21.2%) followed by thriller (18.5%), drama (12%), western (10.4%), action (10.05%), comedy (8.63%) and documentary (7.57%).

Table 3 shows the accuracy, precision and recall corresponding to each of the classification algorithms applied to the dataset.

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fp} + \text{tn} + \text{fn}} \tag{5}$$

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \tag{6}$$

$$\text{Recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \tag{7}$$

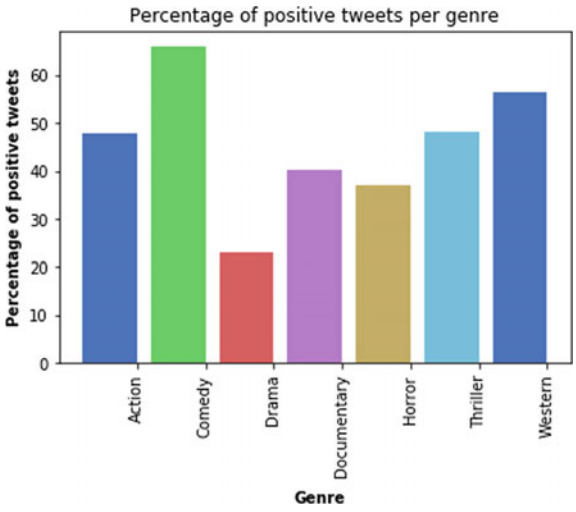

Fig. 3 Percentage of positive tweets per genre

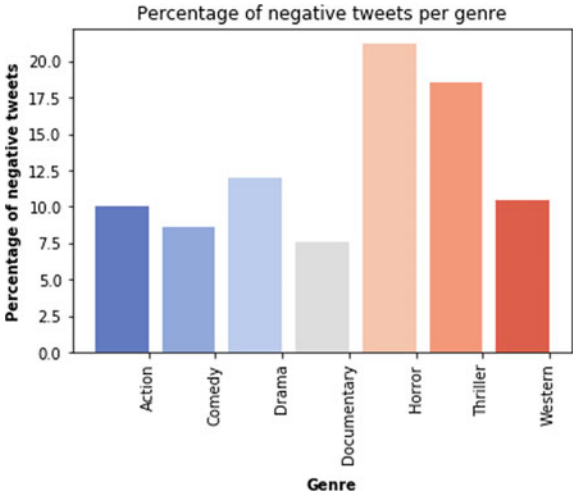

Fig. 4 Percentage of negative tweets per genre

Table 3 Percentage accuracy, precision and recall of the classification algorithms

	Algorithm	Accuracy (%)	Precision (%)	Recall (%)
1	One-versus-rest classifier + logistic regression	74	68.07	41.53
2	One-versus-rest classifier + SVC	83.67	51.16	3834
3	Binary relevance + Gaussian NB	85.33	94.9	79.72
4	Label powerset + logistic regression	80.83	77.14	40.8

where tp stands for true positive, tn is true negative, fn is false negative and fp is false positive. The terms positive and negative suggest the classifier's prediction and the terms true and false allude to whether that prediction belongs to external judgment or observation.

Figure 5 displays the accuracy for each algorithm applied. We can see that the highest accuracy is obtained by using binary relevance plus Gaussian NB algorithm followed by one-versus-rest + SVC, label powerset + logistic regression and One-versus-rest + logistic regression.

6 Conclusion

In this work, we have done classification of movies into various genres. Further, sentiment analysis of the tweets is done using TextBlob. The Movie dataset is used for the experimentation work. We have used different supervised learning algorithms on the dataset and got the best accuracy using binary relevance + Gaussian NB (85.33%). The comedy genre was observed to have the highest percentage of positive tweets, whereas the horror genre was observed to have the highest percentage of negative

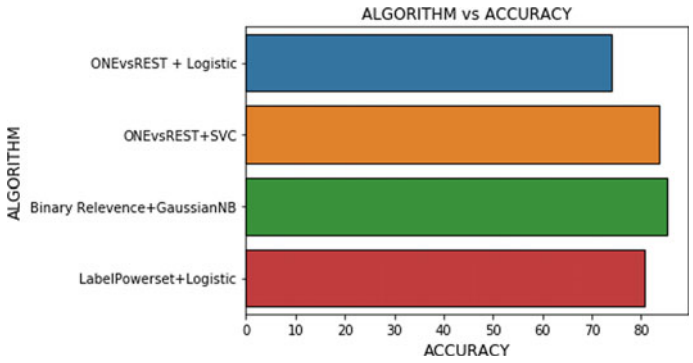

Fig. 5 Comparison of algorithms

reviews. The trends analysis shows that people like comedy shows/movies more than any other genre. It also shows that people are generally more critical of horror movies/tv shows.

References

1. Dangare, C. S., andApte S. S..: Improved study of heart disease prediction system using data mining classification techniques. Int. J. Comput. Appl. **47**(10), 44–48 (2012)
2. Xu, J.: An extended one-versus-rest support vector machine for multi-label classification. Neurocomputing **74**(17), 3114–3124 (2011)
3. Milgram, J., Cheriet, M., Sabourin, R.: "One against one" or "one against all": which one is better for handwriting recognition with SVMs? (2006)
4. Peng, C.Y.J., Lee, K.L., Ingersoll, G.M.: An introduction to logistic regression analysis and reporting. J. Educ. Res. **96**(1), 3–14 (2002)
5. Szymański, P., &Kajdanowicz, T.: Is a data-driven approach still better than random choice with Naive Bayes classifiers?, In: Asian Conference on Intelligent Information and Database Systems, pp. 792–801, Springer (2017)
6. Cheng, W., Hüllermeier, E.: Combining instance-based learning and logistic regression for multilabel classification. Mach. Learn. **76**, 211–225 (2009)
7. Piryani, R., Madhavi, D., Singh, V.K.: Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. Inf. Process. Manage. **53**(1), 122–150 (2017)
8. Munjal, P., Narula, M., Kumar, S., Banati, H.: Twitter sentiments based suggestive framework to predict trends. J. Stat. Manag. Syst. **21**(4), 685–693 (2018)
9. Kadam, T., Saraf, G., Dewadkar, V., &Chate, P. J.: TV show popularity prediction using sentiment analysis in social network. Int. Res. J. Eng. Technol **4**(11) (2017)
10. Mhaigaswali A., Giri N.: Detailed descriptive and predictive analytics with twitter-based TV ratings (IJCAT), vol. 1, pp. 125–130 (2014)
11. Rahim, M. S., Chowdhury, A. E., Islam, M. A., Islam, M. R.: Mining trailers data from youtube for predicting gross income of movies. In 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 551–554). IEEE (2017)
12. Zubiaga, A., Spina, D., Martínez, R., Fresno, V.: Real-time classification of twitter trends. J. Am. Soc. Inf. Sci. **66**(3), 462–473 (2015)
13. Satyavani, A. V., Raveena, M., Poojitha, B.: Analysis and prediction of television show popularity rating using incremental K-Means Algorithm IJMET, vol. 9, pp. 482–489 (2018)

14. Schmit W., Wubben S.: Predicting ratings for new movie releases from Twitter content. Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), pp. 122–126 (2015)
15. Wang, H., Zhang, H.: Movie genre preference prediction using machine learning for customer-based information. Int. J. Comput. Inform. Eng. **11**, 1329–1336 (2017)
16. Battu, V. et al.: Predicting theLsed on its Synopsis, 32nd Pacific Asia Conference on Language, Information and Computation Hong Kong, pp. 52–62 (2018)
17. Maloof, M. A. (Ed.): Machine learning and data mining for computer security: methods and applications. Springer Science & Business Media (2006)
18. Khan, R., Urolagin, S.: Airline sentiment visualization, consumer loyalty measurement and prediction using Twitter data. Int. J. Adv. Comput. Sci. Appl. **9**(6), 380–388 (2018)
19. Bhardwaj, P., Gautam, S., Pahwa, P.: A novel approach to analyze the sentiments of tweets related to TripAdvisor. J. Inf. Optim. Sci. **39**(2), 591–605 (2018)
20. Jindal, R., Taneja, S.: A lexical-semantics-based method for multi-label text categorization using word net. Int. J. Data Mining Model. Manage. **9**(4), 340–360. Publisher: Inderscience (2017)
21. Banik, R.: The Movies Dataset, (Version 7), [Metadata on over 45,000 movies. 26 million ratings fromver 270,000 users.]. Retrieved from https://www.kaggle.com/rounakbanik/the-movies-dataset/metadata [Last Accessed: 15 October 2019] (2017)