Genre Classification of Movie Data Set using Sentiment Analysis

Ashwini M Joshi Assistant Professer  
Department of Computer SciencePES University*,* BSK III Stage Bangalore-560085, India ashwinimjoshi@pes.edu

Shreyas Nitin Pujari   
Department of Computer Science  
PES University*,* BSK III Stage  
Bangalore-560085, Indiapujari.shreyas@gmail.com

Mayank Rao  
Department of Computer Science  
PES University*,* BSK III Stage  
Bangalore-560085, India mayankrao16@gmail.com

Prajna Nagaraj  
Department of Computer Science  
PES University*,* BSK III Stage  
Bangalore-560085, India prajnanag@gmail.com

*Abstract*

*This project aims to classify a given movie into its respective genre by using IMDb movie data set, with positive and negative sentiments included. Using the movie plots generated, the most possible genre is predicted. Sentiment analysis is done with the Machine learning technique RNN using LSTM. Further, genre prediction is done using Logistic Regression and ANN.*

*Keywords – Sentiment Analysis, Recursive Neural Net, Naïve Bayes, Logistic Regression, ANN, Tokenizer.*

# Introduction

In today’s world movies are a very important part of our day to day lives, so before watching the movie, movie reviews plays a vital role in deciding whether or not to watch the movie. The kind of movie also depends on what genre the movie watcher is interested in watching.

In this particular project we focus on classifying the genre based on the plot summary of any given movie. This is done to ensure that the viewer can enjoy the movie according to his/her preference.

For example:- suppose a viewer wants to watch a movie, like The Avengers, the movie clearly classifies into a sci-fi, action genre. Our model tries to classify the same for the viewer. Just knowing the genre is may not be sufficient as the viewer may also be interested to know whether the movie is worth their time. And this is where reviews of the movie come into picture. Our model tries to classify the reviews into two sentiments namely- positive and negative.

# Dataset

## Selecting a Dataset

The dataset is from Kaggle under a title of Movie Reviews IMDb (Internet Movie Database).

This dataset is also present within Keras and was used for the sentiment analysis part of the project.

The dataset used for genre prediction is TMDb (The Movie Database).

## Pre processing Data

The data had to be highly preprocessed before use, consider the fact that the Neural Network only accepts integers and converted vectors, we neglected to use Word2Vec here as the vector orientations of the word have no basis on their similarity, instead first relying on removal of stop words, punctuation and lowering of text.

Therefore, Keras tokenizer was used to convert each sentence into an integer value. This was passed alongside a sentiment value. 1 for positive, and 0 for negative.

A mapping was created to assist in this process, from integer set to text and back.

Further more the genre column of the TMDb dataset had to be reformatted into a json format consisting only of the genres each movie was categorized into, this would aid in the testing phase, where the predication and the correct label are compared. Some movies did not have genres mentioned in the dataset, these were considered as Nan and those entries were removed from the dataset.

# Related Work

The concept for the project requires a lot of pre processing and training data as without the required material it is quite possible to mis classify a lot of data. Simple methods such as detecting negations etc will not work as it is impossible to keep track of every kind of negation as well as the order of the negation.

So two methods were considered, Machine Learning approach and the Probabilistic approach.

With regard to the Machine Learning approach, the effort of detecting the quantity of the text is reduced as the neural network learns from the given set of sentences to recognize the patterns present in everything including the count of negations.

For probabilistic approach, keeping track of the amount of occurrences of data or the relative probability with which a word occurs within a context and using that to classify the document is also a possibility.

As a result, the quality of data from these models is fairly low, as without some supporting model, even these classifiers end up becoming a hit or miss.

Polarity shift is addressed in term counting methods by reversing the orientation of polarity shifted words and then taking the summation of orientation score of each word

# The Experiment

## The high level model of the experiment

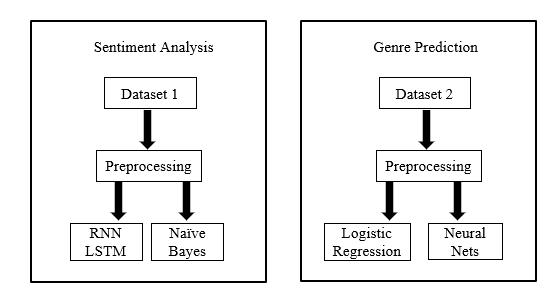


Fig. 1. Sentiment Classification using RNN and LSTM

Genre prediction using Logistic regression and Neural Nets

## Reccurent Neural Network

* Sigmoid activation function is used as we’re trying to predict if the review has a positive or negative sentiment. Therefore, the output layer will be just a single unit, which will be activated by the sigmoid function.
* LSTM is used for the sole reason that it can processes sequences of data rather than individual pieces of text. The memory cells along with the input and output gates provide a fine tuned model. This also means the network can remember it’s state and hence has a better accuracy.
* Binary cross entropy is a perfect loss function as it is being used between only two labels, positive and negative.
* The sequences are padded to some mean length over the length of all reviews. Long ones are shortened, and short ones are elongated. However it is to be noted that short reviews being padded with 0s is a loss in accuracy, as well as shortening the longer reviews as valuable information is lost
* The neural network is assigned to train over a batch size of 64 and passes the aforementioned tokenized sentence to the network.
* Trained over 3 epochs, an accuracy of 86.406% is realized.

## Naïve Bayes

* Gaussian NB is used mainly since we have only two possible classes.
* The same word embeddings are used as RNN since they both consider the same parameters.

## Logistic Regression using ANN

## Some Common Mistakes

* Gaussian NB is required to be used as Multinomial NB etc. work only on multiclass datasets.
* While an SVM can separate the points, the number of outliers due to similarity of the emotions in reviews can lead to a very narrow hyperplane leading to loss in accuracy.
* While converting and using the Tokenizer with the RNN, the vocabulary will be one more than the length of the entire vocabulary.
* The shape of the texts as to be padded to approximately the mean of the size of reviews. Padding too many zeros to a small review will decrease accuracy to a large amount.
* During genre classification, the plot summaries which have a lot of words related to action words, they will be classified into some other genre where the word may also fit.

# Results

|  |  |  |
| --- | --- | --- |
| Dataset | Machine Learning Method | Accuracy |
| IMDb |  |  |
| IMDb |  |  |
| TMDb |  |  |
| TMDb |  |  |

# Conclusion and Future Work

In this paper uses methods of Tokenization, and One hot encoding are used to, identify the most commonly used words and commonly occurring genres in the dataset.

It is observed the accuracy of training and testing can be significantly improved, by using methods of SVM or Word2vec and scaling down on the number of genres being used for classification.

Taking the top ten genres will improve the prediction rates and can be worked on in the future.

##### References

The following data was briefly used during the study of methods to be used for this project.

None of these are singly considered for our method, we wanted to try something different rather than the usual given status quo and decided to borrow pieces from each of these papers.

Some of these papers only served to provide inspiration for other methods or discarded methods, but they helped immensely in the long run.

[1] T. Wilson, J. Weibe, and P. Hoffman on Recognizing Contextual Polarity: An exploration of features for phrase level sentiment analysis, 3rd ed, vol. 35, Computational Linguistics, 2009, pp. 399-433.

[2] Y. Choi and C. Cardie, “Learning with compositional semantics as structural for subsentential sentiment analysis,” in Proc. 2008 Emirical Methods In Natural Language Processing Conf., pp. 793-801.

[3] P. Turney, “Thumbs Up or Thumbs Down? Semantic orientation applied to unsupervised classification of reviews,” in Proc. 2002 Annual Meeting of the Association for Computational Linguistics.

[4] P. Turney and M. L. Littman, “Measuring praise and criticism: Interference of semantic orientation from association,” ACM Transactions on Information Systems (TOIS), vol. 21, 2003, pp. 315- 346.

[5] S. Kim and E. Hovey, “Determining the setiment opinions,” in Proc. 2004 COLING Conference.

[6] A. Kennedy and D. Inkpen, “Sentiment classification of movie revies using valence shifters,”in Computational Intelligence, 2006, vol. 22, pp. 110-125.

[7] B. Pang, L. Lee and S. Vathyanathan, “Thumbs up? Sentiment classification using machine learning techniques” in Proc. 2002 Empirical Methods In Natural Language Processing Conf., pp. 79-89.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**