Movie Genre Prediction and Sentiment Analysis Using Natural Language Processing

DR. VINAYAK ASHOK BHARADI[1], PRATIKSHA SHAILENDRA POSHE[2], GUDIYA ASHOK YADAV[3], RUTUJA SHANTARAM NANDIWADEKAR[4]

[1] *Associate Professor, HOD-IT Department, Finolex Academy of Management and Technology, Mumbai University.*

[2, 3, 4] *Finolex Academy of Management and Technology, Mumbai University.*

Abstract- ***In our daily life we always need some entertainment to relieve our stress. It is great means of spending some quality time with our family members as well. Watching a movie is great source of such entertainment. This is the reason movie have become a large business with good investment. Nowadays Movies are also released on online platform so the viewers' can watch the movie in the comfort of their home. As sometimes theater tickets prizes are very high and due to busy life; people prefer to watch movie based on specific genre to save their times as well as money. At this time, movie genre prediction system plays great role in recommending movies based on viewers' interest.***

In our project, we are using supervised learning approach for prediction of movie genre based on movie subtitles. For execution purpose we are using Google Collab notebook as well as Flair Framework which provides state of the art Natural Language Processing models. The data we are collecting is in raw format so it is necessary to pre-process the data to achieve better accuracy for our classification model. To achieve this, stop word removal methods, Named Entity Recognition (NER) model and other methods are used to extract relevant data. After pre-processing, the dataset will be passed to Artificial Neural Network for genre prediction. Drama, Action, Comedy, Musical, War, Crime, Western, Romance and Horror are the eight categories in which movie transcript will be classified.

Indexed Terms- ***Supervised Learning, Artificial Neural Network (ANN), Flair framework, Pre-processing***

I. INTRODUCTION

Entertainment is necessary part our daily lives. It brings happiness and refreshment in our lives. People do many things to entertain themselves as playing games, going on a picnic, listening to music, watching movies and many more. Many people prefer to enjoy movies with their families. They select movies based on categories of their interest and popularity rating, whether it got positive review or not. As in this super busy and fast track life people don't have much time to surf and select movies of their choice; hence they check out various apps and websites which recommend movies of specific genre. Natural Language Processing is great technology with which we can do such tasks efficiently. It is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. (Natural Language Processing, 2020) Combining above technology with deep learning approach will give more promising results than following traditional approach.

In our task we are combining NLP and Deep learning approach to build a model which predict genres of movies such as drama, romance, horror and action and also analyze their popularity based on live tweets data on Twitter social networking website. We will use Python Programming along with Flair framework.

II. LITERATURE SURVEY

"Semantic Video Classification Based on Subtitles and Domain Terminologies" [1] proposed an algorithm which is based on the WordNet lexical database and the WordNet domain and applies Natural

Language Processing techniques on subtitles. For data pre-processing Mark Hepple's POS tagger is applied and stop words removal method from NLP are used. Then the annotated text is passed to Text Rank algorithm for keyword extraction. In order to retrieve correct contextual meaning of a word and to improve prediction accuracy Word Sense Disambiguation method is used. Subsequently, the WordNet domains that correspond to the correct word senses are identified. The final step assigns category labels to the video content based on the extracted domains. After experimental analysis it is found that the proposed system is very effective.

"Sentiment Analysis of Movies Using Classification Technique to Predict Their Genre Class" [2] proposed a supervised approach for classification based on subtitles. For pre-processing tokenizer, stop Word Handler and stemming techniques are used. Weka Tool is used for handling above techniques. For sentiment analysis purpose comparison-based analysis is done for two classifiers i.e., Support Vector Machine and Random Forest and classification accuracy is analyzed using Percentage Split and Cross Validation testing method. After analyzing results they found that SVM has been the better classifier, providing the best accuracy of 83.33% compare to Random Forest with best accuracy of 75%.

"Neural Network Approach for Text Classification using Relevance Factor as Term Weighing Method**"** [3] proposed Back Propagation Neural Network algorithm for text classification. It is a multilayer feed forward neural network. For Pre-processing, after Stop Word removal Porter Stemmer algorithm is used for stemming. Two supervised term weighing methods which are Binary Term Frequency (tf) and Binary Term Frequency with Relevance factor (tf.rf) are compared. This involves assigning each term a weight indicating the relative importance of the term in a document. Experimental analysis concludes that (tf.rf) performs better than (tf). The performance is evaluated by F-measure term. Without using relevance factor value of f measure is 0.72 while using (tf.rf) gives f measure=0.92. Also, it requires high pre-processing time.

"Real Time Twitter Sentiment Analysis using Natural Language Processing" [4], this paper reports on the design of a data analysis, extracting vast number of tweets. Naïve Bayes machine learning algorithm was used. This paper includes that a program is made to analyze the nature of tweet on a particular topic. In this paper it is mentioned that each tweet extracted classified based on its sentiment whether it is a positive or negative and data were collected on movie reviews which were on IMDB Website. The result from this model was tested using various testing metrics. Moreover, the model demonstrates strong performance on mining texts extracted directly from Twitter. According to this paper, to develop a Machine Learning Model by training the model to categorize the tweets based on sentiment of the tweet and make the model as accurate as possible, first the user will give input i.e., the keyword for extracting the tweets and then the extracted tweets will be categorized by the Machine Learning Model which will be either positive or negative tweet and then the output will be displayed in graphical manner for better understanding of the results.

"A Study on Sentiment Analysis Techniques of Twitter Data" [5], this paper explores the various sentiment analysis applied to Twitter data and their outcomes. The main objective of this paper is to study the existing sentiment analysis methods of Twitter data and provide theoretical comparisons of the state-of-art approaches. The paper is organized as follows: the first two subsequent sections comment on the definitions, motivations, and classification techniques used in sentiment analysis. A number of document level sentiment analysis approaches and sentence-level sentiment analysis approaches are also expressed. Various sentiment-analysis approaches used for Twitter are described including supervised, unsupervised, lexicon, and hybrid approached. Finally, discussions and comparisons of the latter are highlighted.

"Twitter Sentiment Analysis" [6] proposed a machine-based learning approach which is more accurate for analyzing a sentiment; together with natural language processing techniques will be used. This paper reports on the design of a sentiment analysis, extracting a vast amount of tweets. Prototyping is used in this development. Results classify customers' perspective via tweets into positive and negative, which is represented in a pie chart and html page. However, the

program has planned to develop on a web application system, but due to limitation of Django which can be worked on a Linux server or LAMP.

"Sentiment analysis of twitter data" [7], this technical paper shows the application of sentimental analysis and how to connect to Twitter and run sentimental analysis queries. In this paper, social network analysis and its importance are discussed. The text blob python library is used for text processing. Corpora text set is used.

This technical paper reports the implementation of the Twitter sentiment analysis, by utilizing the APIs provided by Twitter itself. There are great works and tools focusing on text mining on social networks. In this, the wealth of available libraries has been used. Two fundamentals approaches are used to extract text

"Predicting the Genre and Rating of a Movie Based on its Synopsis" [8] performed experiments using deep learning models including Convolutional Neural network and Recurrent Neural network with character, word, and sentence level embedding's for inputs. Compared analysis with traditional approaches like SVM and Random Forest showed that deep learning methods performed well than traditional approaches.

III. PROPOSED SYSTEM

1) Problem Statement

In our project, we have proposed to do movie genre prediction and popularity rating using Natural Language Processing. The movie genre such as Drama, Romance, Action and Horror will be automatically detected using transcript of movie subtitle. We are using supervised learning using Neural Network approach to predict genre of a movie. Further the sentiment of a particular movie will be generated by the tweet analysis on live tweet data. The python programming along with Flair framework, NLTK, Tweepy and Keras will be used for the implementation.

2) Data Collection

To build our classification model we need dataset. As we are processing subtitles of movie, we need to store subtitles of a particular movie in a single text file so that processing becomes easier. For this purpose above required data is collected from popular video streaming website YouTube. We have the extracted the movie transcript from video of a movie which is streamed on YouTube.

3) Data Preprocessing

The extracted data is in raw form which is not appropriate for prediction and analysis. We have to eliminate inappropriate data in order to achieve better accuracy of classification model.

Tokenization is necessary and basic step in data preprocessing. As it divide sentence into words and another tokens such as punctuation marks; it becomes easy to perform further preprocessing steps. NLTK library provides tokenizer which tokenizes the data. When we pass the sentence, the text is automatically tokenized using this lightweight library.

After tokenization stop word removal is necessary so that more focus can be given to those words which define the meaning of the text. We will use NLTK library for stop word removal. NLTK has list of stop words stored in 16 different languages. For tagging, we will use pre-trained model for Named Entity Recognition (NER). This model was trained over the English CoNLL-03 task with accuracy 93.03.

Word embedding or vectorization is a method to map words or phrases from vocabulary to a corresponding vector of real number which is used to find word predictions, word similarity or semantic. Flair framework supports many categories of word embedding's such as Glove, Word2Vec, ELMo, Document embedding's, Flair embedding's etc. In Flair, combinations of word embedding are also possible. As we are dealing with movie sentiment we will use combination of embedding's. For combination of embedding Flairs Stackedembeddings class is used.

4) Movie Genre Prediction

We will classify movie transcripts into eight genres as Drama, Action, Comedy, Musical, War, Crime, Western, Romance and Horror. For this prediction we will use Artificial Neural Network (ANN).

5) Sentiment Analysis

For sentiment analysis, we will use live tweet data provided by twitter social networking website. We will scrape live tweets on a movie by using tweepy

library. After scraping tweets data, relevant tags will be extracted using Named Entity Recognition model (NER) in Flair. After extracting relevant tags, sentiment of data is obtained by passing data to Flair's Pre-trained Text Classifier 'en-sentiment' which is a sentiment analysis model based on IMDB dataset training and an offensive language detection model.

IV. FLOW DIAGRAM

1) Sentiment Analysis

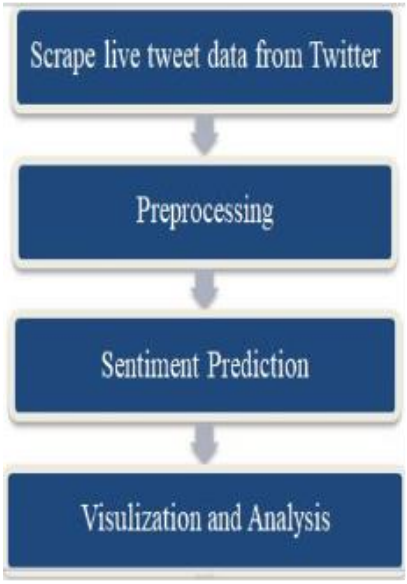


Fig.1. Flow diagram for sentiment analysis

2) Movie Genre Prediction

Fig.2. Flow diagram for movie genre prediction

V. EVALUATION MODELS

1. Flair Framework

Flair is a simple natural language processing (NLP) library developed and open-sourced by Zalando Research. PyTorch used to build Flair framework, Flair is one of the best deep learning frameworks. Hence, we are going to use Flair Framework to implement our project related work. It helps you to apply state-of-the-art NLP models to the words.

Following are some state-of-the-art natural language processing models which we are going to apply to our text:

1. Name-Entity Recognition (NER): It is use to acknowledge whether a word signify a person, location or names and organization in the text or not. This model was trained over the English CoNL-03 task. We are going to use NER model to identify the above-mentioned categories in the text of our provided dataset.
2. Parts-of-Speech Tagging (PoS): It will help to claim "which part of speech" (e.g., noun, pronoun) does all the words in the given text belongs. Flair framework support multiple models i.e., Fast English Model, Multilingual Model, German Model so, as per our project requirement we will use these models in our task. Flair also supports Multi-Tagging and Stack-Tagging we are going to use this for better accuracy of the output.
3. Embedding: Flair framework's this method is used to map words or phrases from vocabulary to a corresponding vector (real number). There are some common embedding types are there such as Classic Word Embedding, Stacked Embedding and Flair Embedding Flair framework supports many categories of word embedding i.e. Glove Word2Vec, ELMo. The combination of word embedding is also possible in Flair framework.

2. Artificial Neural Network

Artificial neural network is a multi-layer fully connected (Multiple nodes are connected to each other and they are hidden) neural network. ANN is used to applied for learning real-valued, vector-valued and discrete-valued functions which contains problems like speech recognition, interpretation of visual scenes. Artificial Neural networks has two layers one

is called as input layer and the other one is output layer and it has some hidden layer also.

ANN is a mathematical model which is neurally implemented. It has large number of interconnected neurons which performs all the operations. The arrangements of these neurons are very important and essential in Artificial Neural Network Neurons is also used to store the information basically called as weighted linkage of neuron. It has ability to recall, generalize and learn from given data.

VI. RESULTS & DISCUSSION

For Movie Sentiment Analysis

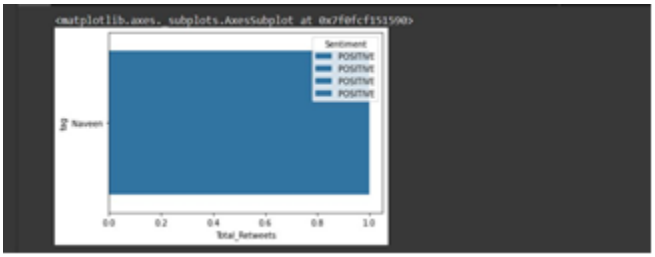
Movie name is passed as input to twitter search query.

All the downloaded tweets are stored into the .csv file

From Flair library Text classifier named en-sentiment is downloaded

After all the processing and predicting overall sentiment column based on Average polarity of the tag is created.

Graph is plotted showing sentiment of an inputted movie.



For Movie Genre Prediction

Movie Subtitle file is passed to model as input and genre is predicted

CONCLUSION

In the course of our study, we understood that for movie genre prediction from subtitles, overall context of subtitle transcript is necessary. We have studied framework and approaches which are useful for genre prediction and sentiment analysis of a movie. Flair framework is a state-of-the-art NLP library which provides many functionality and libraries for pre-processing and classification purpose. It gives better result and improved accuracy than any other framework. As here context is necessary for prediction, Artificial Neural Network (ANN) including multi-layer fully connected neural nets consists of an input layer, multiple hidden layer and output layers. Hence, we have used these approaches to work on prediction of movie genre and popularity analysis.

REFERENCES

[1] Polyxeni Katsiouli, Vassileios Tsetsos, Stathes Hadjiefthymiades, "Semantic video classification based on subtitles and domain terminologies,".

[2] Pankaj Kumar, "Semantic video classification based on subtitles and domain terminologies Sentiment Analysis of Movies Using Classification Technique to Predict Their Genre Class," School of Computer Science and Engineering Lovely Professional University Phagwara, Punjab (India) Month April, Year 2017.

[3] Anuradha Patra Barkatulallah and Divakar Singh, "Neural network approach for text classification using relevance factor as term weighing method" International Journal of Computer Applications (0975 – 8887) Volume 68– No.17, April 2013.

[4] Anupama B S, Rakshith D B, Rahul Kumar M, Navaneeth M,"Real time twitter sentiment analysis using natural language processing"

[5] Hamid Bagheri, Md Johirul Islam, "A study on sentiment analysis techniques of twitter data"

[6] Aliza Sarlan, Chayanit Nadam, Shuib Basri,"Twitter sentiment analysis"

[7] Abdullah Alsaeedi, Mohammad Zubair Khan, "Sentiment analysis of twitter data"

[8] Varshit Battu, Vishal Batchu, Rama Rohit Reddy, Murali Krishna Reddy, Radhika Mamidi, "Predicting the Genre and Rating of a Movie Based on its Synopsis", International Institute of Information Technology Hyderabad.

[9] Akbik, Alan and Blythe, Duncan and Vollgraf, Roland, "Contextual string embeddings FOR sequence labeling", {COLING} 2018, 27th International Conference on Computational Linguistics, 1638--1649, 2018.

[10] Sharoon Saxena, "Introduction to Flair For NLP: Simple yet powerful state of the art NLP Library" ,https://www.analyticsvidhya.com/blog/2019/02/flair-nlp-library-python/

[11] "introduction-to-artificial-neutralnetworks", https://www.geeksforgeeks.org/introduction-to-artificial-neutral-networks

[12] Bilal Tahir, "Twitter pulse checker: an interactive colab notebook for data sciencing on twitter "https://towardsdatascience.com/twitter-pulse-checker-an-interactive-colab-notebook-for-data-sciencing-on-twitter-76a27ec8526f