JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Twitter User Personalization Using NLP

[1]Smt. N V Ramya Devi Kotla, [2]V Jaya Lakshmi, [2]Y Nehasri, [2]G Navya

[1] Assistant Professor of SRKR Engineering College, Dept of IT, Bhimavaram-534204, Andhra Pradesh, India
[2]Students of SRKR Engineering College, Dept of IT, Bhimavaram-534204, Andhra Pradesh, India
jayavarra1@gmail.com

Abstract **-** One of the major components of social media is to connect people having similar opinions and interests. Also, the pioneering applications of social media that has an ability to dictate a person's day to day life and a platform for people to share their opinion is Twitter. In twitter recommendations are based on hashtags, but it may not be the correct way of recommending tweets. The main objective of this work is to develop the recommendation system based on topic and sentiment. This will make group of people to share their opinions among them. For this work, we collected various datasets from Kaggle website and merged to form a new dataset. We implemented this work using LSTM model, Logistic Regression Model, SVM model. The accuracy obtained by LSTM model is 80%, which is more than Logistic Regression and SVM model. Among those methods LSTM model shows great accuracy when compared to other models. So, this system is successfully identifying the sentiment and topic of a tweet using LSTM model.

Introduction

Twitter is a social media platform that allows people to share their opinion and even as a news outlet of what's happening around the world. It connects the people among the world who hold a very diverse range of opinions and values. The topics that are opinionated on by the people of twitter are very diverse and vast in nature. Sentiment analysis is widely used by so many marketing companies to know their opinions and understand the market trends.

A combination of sentiment analysis and classification of the topics based on topics will classify the tweets as n number of groups. This will allow us to analyze the sentiment of the tweet and what topic it was opinionated on. From this analysis we will understand what topic are most discussed and how the internet is biased to which opinion. For this objective, there is a need for a social science tool in order to understand the importance of public sentiments and topic of the tweet. This information will help the marketing organizations to find the sentiment of the people.

This system uses core concepts of machine learning and deep learning to provide accurate or near to accurate predictions and classification of the given tweets based on Natural Language processing which is one of the complex practices of machine learning. This system aims to recommend tweets based on sentiment and topic. Specifically, the project will classify tweets as positive, negative, and then recommend tweets related to a specific topic with a sentiment. This technology has the potential to be used in several applications, including identifying public sentiment towards brands or products, filtering out irrelevant or negative tweets, and identifying trends and patterns in social media data to classify tweets into different topics. Overall, a topic-based tweet recommendation system has the potential to improve the user experience by presenting the most relevant and interesting tweets to users.

Technologies used: NLP, Machine Learning using Python.

Python libraries and Packages: NumPy, pandas, matplotlib, sklearn, Keras, Cosine Similarity.

I. LITERATURE SURVEY (RELATED WORK)

V. K. Singh, et.al [1] stated that By including a sentiment classifier in the recommendation process, they proposed an alternative strategy to a hybrid recommender system that enhances the outcomes of collaborative filtering. Collaborative filtering served as the first level of filtering for their experiment with the movie reviews dataset, and the sentiment classifier served as the second level of filtering. They put this into practice using Nave Classifier models. The outcomes lead to a more targeted and superior movie recommendation.

H. John, et.al [2] looked at a variety of profiling and recommendation techniques to show the potential for follow-up recommendations that are both effective and efficient. They compared accuracy using large Twitter datasets, logistic regression, and SVM techniques. When the amount of data is small, the results demonstrate that Logistic Regression performs better than SVM model. They concluded that small data works better for logistic regression and large data increases the training time for SVM.

J. Chen, et.al [3] claimed that relationship-based algorithms are more effective at locating known contacts. All algorithms were successful at growing users' friend lists, according to their research. While algorithms based on user-created content were better at finding new friends, algorithms based on social network information were able to produce recommendations that users were more likely to accept and locate more familiar contacts for them. When they used the SONAR model and the IBM reviews dataset in their experiments, the accuracy was 80%.

Garcia, et.al [4] showed that a novel, interesting, and personalized method of connection recommendation could be to weight the impact of features in a personalized way. They employed an SVM model and a dataset of mobile reviews. The results indicate that using two features instead of one

results in a small but significant improvement in performance, and they suggest that adding more features will lead to even greater improvements in predictions.

Trstenjak, et.al [5] suggested combining the KNN algorithm with the TF-IDF framework and method for text classification. They took a document with several tweets in it. With a few minor adjustments to their implementation, the KNN algorithm and TF-IDF method have been shown to be a good match. The framework enables the embedded classification algorithm to be updated and enhanced.

II. METHODOLOGY

OBJECTIVE -The main objective is to classify tweets into positive or negative tweets and recommend tweets based on sentiment and topic. Predicting sentiment and recommending tweets by considering few features out of all from dataset. To create a model that can predict more accurate results. To analyze necessary features required to be used in real-time applications.

DATASET -The dataset we used is taken from multiple datasets that are present on Kaggle and we merge them into a single dataset. The datasets that are taken from Kaggle website are Movie reviews dataset, Mobile Reviews Dataset, Food reviews Dataset. These datasets are combined using Python Libraries to form a new dataset that is used for this Project.

DATASET DESCRIPTION -The dataset consists of three columns [Text, Target, Topic]. Text Column Contains tweets from various users. Target Column contains two values 0 and 1 .0 represents that the tweet is Negative and 1 Represents that the tweet is Positive. Topic Columns Consists of several values representing in which topic they are tweeted about.

DATA PRE-PROCESSING -It includes the following steps.

Spaces strip: Removing of leading and trailing white spaces.

Removing punctuations: While pre-processing the text data punctuations may not be necessary to find emotions of a tweet.

Therefore, removing all instances of it will help us reducing the size of training data.

Ex: @, $ etc.

Lower case conversion: Transform tweets from upper case to lower case. This helps in reducing the comparison of two same words but having upper case and another one is in lower case.

Stop words removal: commonly occurring words should be removed from text data.

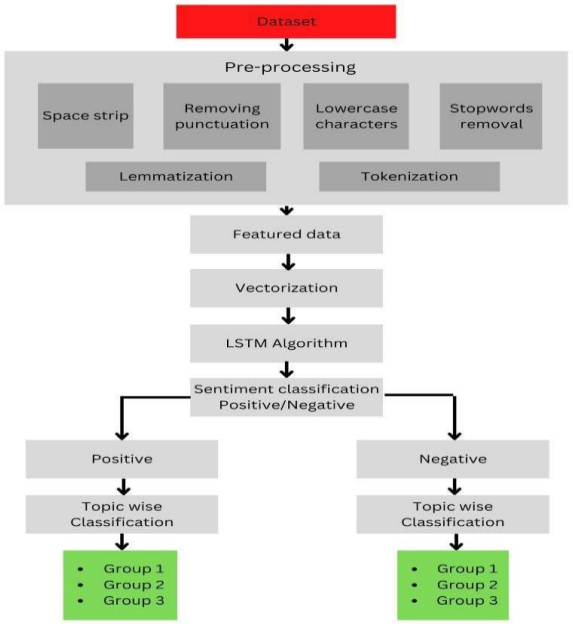
stop words= [‘a’, ’an’, ’the’, ’are’, ’to’, ’and’] and so on…

Removal of URLs: Links found in the text are removed.

Applying lemmatize: It converts the word into its root word, rather than just stripping the suffixes.

Ex: Caring: care

Tokenization: This separates the sentence into tokens



III. MODELS USED

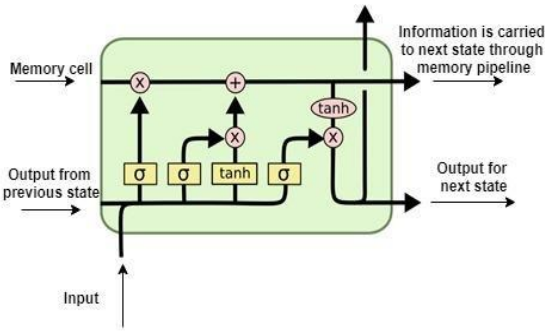
Here we want to use different models and compare the model’s accuracy with each other to decide which model gives best results. We are using different models like LSTM model, Logistic Regression, SVM model and comparing these models with accuracy.

LSTM:

LSTM (Long Short-Term Memory) and RNN are two deep learning models that we are using in our project to categories user emotions in tweets. (Recurrent Neural Network). For training and prediction purposes, the featured vectors were combined and fed into LSTM and RNN to classify the tweet's emotions into positive and negative sentiments.

The LSTM has demonstrated its effectiveness in classifying emotions with excellent accuracy. The cell that stores state information can be updated or deleted by the LSTM, and gates are in control of these operations. Information can move through networks thanks to gates.

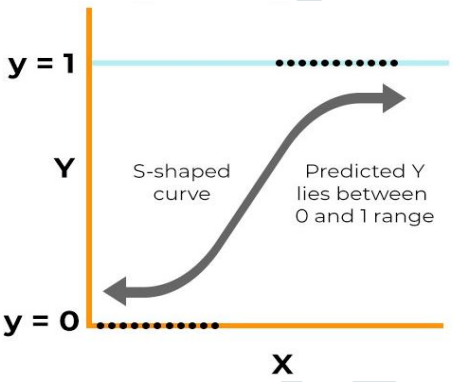
The choice of whether to accept or forget the features is made by the LSTM's forget layer. Depending on the sentiment of the tweets, the new emotion word must be accepted or rejected in each cell state, it depicts the LSTM model's architecture.



LOGISTIC REGRESSION:

Logistic regression is a popular method for binary classification problems, where the goal is to predict one of two possible classes - in this case, positive or negative sentiment. In the context of this project, logistic regression can be used to predict the sentiment of a given tweet, based on a set of input features.

Once the input features have been identified, logistic regression can be used to train a binary classifier that predicts the sentiment of a tweet. During training, the model learns a set of weights that map the input features to the probability of the tweet being positive or negative. During prediction, the model takes in the input features of a new tweet and computes the predicted probability of the tweet being positive or negative. This probability can be thresholded at a certain value to obtain a binary classification of the tweet's sentiment.



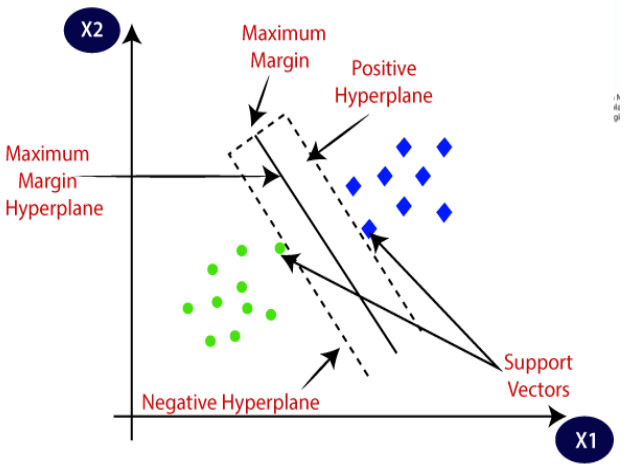
SVM (Support Vector Machine):

We are also testing this project using SVM to get better results when compared to LSTM or Logistic Regression. To use SVM for sentiment analysis, we first need to extract a set of features from the tweets. These features could include information about the tweet text, such as the presence of certain keywords or phrases, the length of the tweet, and so on. Once the features have been extracted, they can be used to train an SVM classifier to predict the sentiment of a given tweet.

During training, the SVM learns a hyperplane that separates the positive and negative tweets in feature space, with the goal of maximizing the margin between the two classes. This hyperplane can then be used to predict the sentiment of a new tweet by projecting it onto the feature space and classifying it based on which side of the hyperplane it falls on it.

One advantage of using SVM for sentiment analysis is that it is a powerful and flexible algorithm that can work well with high-dimensional feature spaces. Overall, SVM is a useful and versatile algorithm that can be applied to a wide range of sentiment analysis tasks, including those involving tweets.

SVM and logistic regression. Among all those models LSTM gives better accurate results. And For topic Classification we used Cosine Similarity metric to recommend tweets. So, we concluded that this system is successful in classifying tweet



IV. IMPLEMENTATION & RESULT

After pre-processing the data, next step is to train the model by splitting data into training and testing datasets, in our case we are taking 80% into training data and 20% into testing data for model implementation.

Here LSTM first converts the data into numerical form and model is trained to predict whether the tweet is positive or negative.

After classifying the tweet into positive or negative tweet, Then Cosine Similarity comes into action. We make a cluster that belongs to same sentiment and same topic and with the help of cosine similarity metric it finds the similarity score between tweets present in the cluster and the input tweet and based on the similarity score it recommends the tweets. In LSTM we are using SoftMax Activation Function to perform Training. In topic Modelling, a new data frame is created from an original dataset based on topic and resulted sentiment. Then, Similarity Score is calculated between new data frame data and input tweet.

Generally, if similarity score is 0 then it states that there is no similarity between two texts and if similarity score is 1 then is states that there is more similarity between two texts. Based on High similarities tweets are recommended.

The LSTM model we used in this project predicted with about 80% accuracy and we used performance of this model.

V. CONCLUSION

To classify tweets into positive or negative and recommending tweets based on topic, we proposed a system which consists of machine learning and deep learning algorithms like LSTM, Logistic Regression and SVM model. We conducted experiment on the collected datasets from Kaggle websites and merged to form a new dataset. Among them comparison is done and the accuracies obtained by LSTM is 80%, which is more than compared to

VI. FUTURE WORK

In the realm of Twitter user personalization through natural language processing, future work could delve into enhancing user modeling by incorporating diverse data sources like demographics and social network structures. Dynamic adaptation to evolving user preferences over time presents an intriguing avenue for exploration, as does the integration of multimodal analysis techniques for a more holistic

understanding of user-generated content. Contextualizing tweets within broader events and user-specific contexts could lead to more nuanced personalization, while ensuring privacy-preserving methods remain a priority. Additionally, there's a need for robust evaluation metrics beyond traditional accuracy, considering aspects like user satisfaction and algorithmic fairness. Scalable deployment, ethical considerations, and domain-specific tailoring further enrich the landscape of potential research endeavors in this domain.

VII. REFERENCES

[1] Singh VK, Mukherjee M, Mehta GK. "Combining collaborative filtering and sentiment classification for improved movie recommendations". In Multi-disciplinary Trends in Artificial Intelligence: 5th International Workshop, MIWAI 2011, Hyderabad, India, December 7-9, 2017. Proceedings 5 2017 pp. 38-50.

[2] Hannon, J., Bennett, M. and Smyth, B., "Recommending twitter users to follow using content and collaborative filtering approaches". Proceedings of the fourth ACM conference on Recommender systems, pp. 199-206, 2018.

[3] Chen, J., Geyer, W., Dugan, C., Muller, M. and Guy, I., "Make new friends, but keep the old: recommending people on social networking sites". Proceedings of the SIGCHI conference on human factors in computing systems, pp. 201-210, 2019.

[4] Garcia-Gavilanes, R.O.G.G. and Amratian, X., "Weighted content-based methods for recommending connections in online social networks"., pp. 68–71 2018.

[5] Trstenjak, B., Mikac, S. and Donkor, D., "KNN with TF-IDF based framework for text categorization". Procedia Engineering, *69*, pp.1356-1364. 2017.

[6] Shou Zhong, T. and Minlie, H., 2016." Mining microblog user interests based on TextRank with TFIDF factor". The Journal of China Universities of Posts and Telecommunications, 23(5), pp.40-46.