Sentiment Analysis for YouTube Comments in Roman Urdu

Tooba Tehreem  
National University of Computer and Emerging Sciences Islamabad, Pakistan  
Email: i191240@nu.edu.pk

Hira Tahir  
National University of Computer and Emerging Sciences Islamabad, Pakistan  
Email: i192071@nu.edu.pk

Abstract
Sentiment analysis is a vast area in the Machine learning domain. A lot of work is done on datasets and their analysis of the English Language. In Pakistan, a huge amount of data is in roman Urdu language, it is scattered all over the social sites including Twitter, YouTube, Facebook and similar applications. In this study the focus domain of dataset gathering is YouTube comments. The Dataset contains the comments of people over different Pakistani dramas and TV shows. The Dataset contains multi-class classification that is grouped The comments into positive, negative and neutral sentiment. In this Study comparative analysis is done for five supervised learning Algorithms including linear regression, SVM, KNN, Multi layer Perceptron and Naive-Bayes classifier. Accuracy, recall, precision and F-measure are used for measuring performance. Results show that accuracy of SVM is 64 percent, which is better than the rest of the list.

Keywords: Sentiment analysis, Machine learning, Roman urdu language, comparative analysis, SVM

1. Problem statement
As social networks are growing in popularity due to their quick and easy access and their low cost. To analyze the sentiments through User roman comments on YouTube by using roman Urdu.

2. Introduction

• To get sentiment analysis results from comments / texts from Roman Urdu. The authors [8] did their best to include all research done so far in the classification of the Roman Urdu. During this study, they observe that the classification was done in two Or three grades, if it was (Negative, Positive or Neutral). For this purpose, most researchers have used logistic regression, support vector machine and naive Bayes.

• Motivation: Since almost all Pakistanis use Roman Urdu as the language of communication, there is a lot of data on the Internet in Roman Urdu that can be used to analyze feelings.

• Background: YouTube is an area where users can view, report, rate, comment, share, upload, and bookmark on videos and

3. Related work
Much work has been done on the sentiment analysis with YouTube. Benkhelifa, et al. [5] discussed the statement extraction and classification of comments on cooking recipes from Youtube in real time. A real-time system has been proposed that automatically extracts and YouTube are categorized cooking recipes. After collecting the Data, filter the comments and classify the comments based on SVM classifier.

Yu et al. [21] Suggested a method for predicting reviews from on-
line prescription users. About the ingredients of the recipe, recipe instructions and reviews are taken into account. The Multi-class SVM was used to study the reliability of this information. In their study, they found that the journal information provided the most reliable predictions.

Timoney et al. [19] have been doing sentiment analysis on YouTube Videos of the best British songs since 1960. Khan et al. [12] conducted a sentiment analysis of UK film data and Amazon product review data the semi-supervised approach. The Lexical methodology has been combined with machine learning to achieve improvements in their study of sentiment analysis. Uysal [20] used the various function selection methods with supervised classification techniques on YouTube Remarks. [3] described the decision tree sentiment analysis. [17] described the Roman Urdu text concepts. This paper [16] used to describe the emotions and behavior of human beings on social media.

Paper [4] described the Cyprus classification and extraction. The author of [14] shows the improvement in automatic speech recognition. [9] described the health care analysis. These papers [25] [13] described mobile reviews using F1-score applying on Roman Urdu. These [2] [8] [15] used for detection purposes using deep learning. [24] [23] [22] described the game sensation. The paper [1] studied about the emotion of corpus using Roman Urdu. These papers [7] [6] described energy consumption for Android Studio.

Paper [18] described the improvement in energy consumption.

4. Proposed approach

The model proposed is split into four major steps. First one is, the Reviews written in Roman Urdu are gathered in the file and labelled manually into 3 categories, i.e. positive, negative and neutral. Then data is preprocessed. After that, features are selected. Data is divided into training and testing. Different models are used by applying different classifiers, and then the outcomes are inspected and relate. The methodology includes the steps shown in the following figure:

4.1 Dataset

Dataset for this study was downloaded from Kaggle. This is a labeled dataset which contains 2 columns. The 1st Column consists of Comments gathered from YouTube in roman Urdu whereas 2nd Column consists of class labels. This dataset comprises of ternary classifications i.e. all comments are divided into different classes i.e. positive, negative and neutral. The First column is for comments, second column is of sentiment, whereas the Third column is titled as nan, which contain null values so we removed that column from the dataset.

The Dataset contains 14131 unique records. Percentage of neutral comments is 48 percent, positive comments is 29 percent and negative comments is 24 percent.

Below bar graph in fig shows the dataset division with respect to labels.

4.2 Preprocessing

Preprocessing is done in 5 steps as shown in fig:

4.2.1 Stop words List creation

A list of roman Urdu stop words are created manually, which contains around 100 words.

4.2.2 Row Extraction

Each row is extracted containing comment from the data frame one by one in order to apply preprocessing steps on it.

4.2.3 Lower case conversion

Upper case letters are converted into lower case using lower () method.
4.2.4 Tokenization
Each row is split into tokens of words using split () method.

4.2.5 Removal of Stop Words
In this step, stop words are removed from the row and remaining comment is added into a new column of a dataset with name 'text_final' which is appended in the original dataset for further use. It can be clearly seen in following Fig that stop words has been removed and preprocessed text is stored in the third column of the dataset.

4.3 Feature Engineering
The purpose of this step is to convert textual data into numerical values in order to feed them to classifiers. There are many ways to do so, but the method used in this study is ‘TF-IDF’ which stands for Term Frequency- Inverse Document Frequency.

In this method, frequency of each word in the data frame is calculated. A unique number and frequency is also assigned to each word. This is also known as word vectorization. TfidfVectorizer can be easily imported from sklearn library of feature extraction.

```
from sklearn.feature_extraction.text import TfidfVectorizer

The TF-IDF model is fitted on the whole dataset by using the method of TfidfVectorizer.Fit() which takes the preprocessed column of dataset as an input and provide a vocabulary of words with unique integer numbers attached to them as an output.

In this way, the dataset is ready to become as an input of classifiers. Glimpse of frequency of each word is shown in following figure:
```

4.4 Train and Test Split
The Dataset divides into two halves, one use for Training while the other use for Testing. Classification models fit on the training dataset while the test dataset uses for making predictions. Sklearn library use for splitting the data on two halves and training-testing ratio mention as an input for the function.80 percent of dataset provides for training, whereas predictions performed on 20 percent of the remaining dataset. Parameter for the training set to 20 percent.

4.5 Machine Learning Algorithms
In this study, a total of 5 ML algorithms are used which are discussed above.
All models are imported and run using the Sklearn python library. Selecting an appropriate model for dataset is crucial because accuracy in prediction depends very much on the model. As discussed in the introductory session, below classifiers are used for this comparative study.

- Naive Bayes Classifier
- Linear Regression
- Support Vector Machine
- KNN
- Multi-Layer Perceptron

5. Evaluation and Experiments
For evaluation purpose of sentiment analysis confusion metrics is used as an aid which shows the values of true positive, true negative, false positive and false negative parameters. These parameters are further used to calculate accuracy score, precision score, recall score and F1-measure score. Precision and recall combine to generate F1-score.

In our experiment, we imported the evaluation metrics from the sci-kit library and used them in python code. Accuracy of all algorithms is discussed individually later on in this paper.

All classifiers are run 10 times and results are stored in the list. Mean values are calculated for accurate and closest results.

5.1 Naïve-Bayes algorithm
Naive Bayesian classifiers are a family of simple "probabilistic classifiers" based on the application of Bayes’ theorem with strong assumptions about independence between features. They are some of the simplest Bayesian network models, but when combined with the estimation of the core density, they can achieve a higher level of precision.

Confusion metrics for Naive-Bayes.

```
|        |        |        |
|--------|--------|--------|
| Positive| Positive| Positive|
| 127    | 141    | 53     |
| Positive| Negative| Negative|
| 33     | 594    | 72     |
| Negative| Positive| Negative|
| 29     | 187    | 229    |
```

5.2 Linear regression
In statistics, linear regression is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables. The case of an explanatory variable is called simple linear regression. If there are several, the process is called multiple linear regression.

Confusion Matrix of Linear regression.
5.3 Support Vector Machine
In machine learning, supporting vector machines are kept an eye on learning models using associated learning algorithms that examine data for regression analysis and classification.
Confusion Matrix of SVM.

\[
\begin{array}{ccc}
165 & 111 & 45 \\
55 & 560 & 84 \\
51 & 161 & 233 \\
\end{array}
\]

5.4 KNN
K’s Nearest Neighbor (KNN) is a manageable, user friendly, flexible algorithm and one of the best in machine learning. KNN is used in a variety of applications such as image recognition, finance, political science, healthcare, video recognition and handwriting recognition.
Confusion Matrix of KNN.

\[
\begin{array}{ccc}
165 & 111 & 45 \\
55 & 560 & 84 \\
51 & 161 & 233 \\
\end{array}
\]

5.5 MLP
A multilayer perceptron is a kind of artificial neural network with direct feedback. The term MLP is used vague, at times loosely

6. Results:
The Ten-fold cross validation technique is used to evaluate the performance of classifiers. The Dataset is irregularly separated into
training and testing partitions. 90 percent data are used for training and 10 percent data are used for testing.

| Classifier   | Accuracy | Recall | Precision | F1-Score |
|--------------|----------|--------|-----------|----------|
| Naive Bayes  | 63       | 63.13  | 63.14     | 63.13    |
| Linear Regression | 64      | 63.7   | 63.81     | 63.88    |
| SVM          | 64       | 64.8   | 64.8      | 64.8     |
| KNN          | 59       | 59.88  | 59.87     | 59.88    |
| MLP          | 56.9     | 57.3   | 57.3      | 57.3     |

**Figure 13.**

Accuracy graph
Following is the accuracy graph in figure 14.

**Figure 14.** Accuracy Graph

Precision graph
Following is the precision graph in figure 15.

**Figure 15.** Precision Graph

Recall graph
Following is the recall graph in figure 16.

**Figure 16.** Recall Graph

**Figure 17.** F1-score Graph

**7. Conclusion**

In this study 5 supervised learning algorithms are applied to a dataset to predict which classifier performs better on roman-Urdu dataset. Feature extraction technique used for this study is anagram which is also known as a bag of words that contains all words in the form of tokens. Maximum features selected for this study from the feature vector are 3000. Our results show that SVM is better in terms of accuracy as relate to other four algorithms. In future work, we will apply multiple feature extraction ability in order to upgrade the all-around precision of the models.

**References**

[1] Muhammad Umair Arshad, Muhammad Farrukh Bashir, Adil Majeed, Waseem Shazad, and Mirza Omer Beg. Corpus for emotion detection on roman urdu. In 2019 22nd International Multitopic Conference (INMIC), pages 1–6. IEEE, 2019.

[2] Muhammad Asad, Muhammad Asim, Talha Javed, Mirza O Beg, Hasan Muntaka, and Sohail Abbas. Deepdetext: detection of distributed denial of service attacks using deep learning. *The Computer Journal*, 63(7):983–994, 2020.
