Title: Sentiment Analysis of Roman Urdu Reviews

Author(s): Muhammad Abdullah Aish, Muhammad Mubeen, Muhammad Umer Iqbal, Abdullah Qaseem

Affiliation(s): Department of Computer Science, National College of Business Administration and Economics

DOI: https://doi.org/10.32350.umt-air.21.01

History: Received: June 30, 2022, Revised: August 9, 2022, Accepted: September 1, 2022

Citation: M. A. Aish, M. Mubeen, M. U. Iqbal, and A. Qaseem, “Sentiment analysis of roman urdu reviews,” UMT Artif. Intell. Rev., vol. 2, no. 1, pp. 00-00, 2022, doi: https://doi.org/10.32350.umt-air.21.01

Copyright: © The Authors

Licensing: This article is open access and is distributed under the terms of Creative Commons Attribution 4.0 International License

Conflict of Interest: Author(s) declared no conflict of interest
Sentiment Analysis of Roman Urdu Reviews

Muhammad Abdullah Aish*, Muhammad Mubeen, Muhammad Umer Iqbal, Abdullah Qaseem

Department of Computer Science, National College of Business Administration and Economics

Abstract—This study aims to analyze easy access and economic availability of computers, tabs, smartphones, and high-speed internet. Moreover, now a day’s people use web/online medium for their social interaction and business correspondence. According to the current research, people are more prone to post their reviews about any specific entity/product that they use in their daily routine. These reviews are very helpful for both—user and seller. Initially, these reviews were not too many and could easily be analyzed by only giving them a read. The continuous increase in the number of these reviews proves that reviews can be analyzed them and useful patterns can be explored through automated channels. This need leads to a new find in the domain of research which is known as “Sentiment Analysis”. It is the study of people’s opinions, sentiments, attitudes, and emotions, expressed in written discourse. Also, it is said that Sentiment Analysis is a process of categorizing people’s opinions expressed textually, in order to determine, whether the writer’s attitude towards a particular topic or product is positive, negative, or neutral. The current research targets the mining process of sentiments from the reviews of PSL anthems. In this particular research work, five different classification models were used for further text classification of reviews, conducted by using the Rapid Miner Tool. The current study presents the Sentiment Analysis of Roman Urdu reviews on PSL Anthems which are available on YouTube channel’s comment section. These reviews were scraped, pre-processed, and analyzed by using the Naïve Bayes, Gradient Boost Tree, Support Vector Machine, K-Nearest Neighbors, and Artificial Neural Network. The Roman Urdu Sentiment Analysis was performed at 7000 Bi-lingual reviews. Naïve Bayes and Logistic Regression correctly predicted 68.86% reviews. While, ANN achieved 68.86% on testing dataset and 69.71% on the validation of the results.

Index Terms—Deep learning, machine learning, natural language processing, natural language reviews, reviews analysis, sentiment analysis, supervised learning

* Corresponding author: abdullahaish7@gmail.com
I. Introduction

As the use of the internet for social networks expands, the rapid evolution of social connection also increases because of the convenience of computers, smartphones, and digital tablets. Also, broadband has made people utilize web services for their social relations, business resemblance, e-commerce, and to make social contacts [1]. These days, people persistently give their online feedbacks under the review section of the relevant page or channel, that is, comments/reviews about any specific entity/product they have used. As they feel comfortable with online purchasing, they tend to increasingly share their reviews about any particular entity, and also share their experience with other users. Now a days, customer’s reviews have become an important source to understand the feedback and responses towards any specific entity or product [2]. Moreover, few companies were there in the past that carried out a market survey by taking products in their hands and made conclusions about them. Furthermore, this may benefits the organization to improve their quality standards and also it helps users to check the credibility of product by viewing the reviews of previous users [3]. This data collection method is outdated. Now, the product makers give a place to the users to share their experience with the other at the same place where the information of the product is displayed to rate their selling commodity in the form of rating, likes, dislikes, reviews and comments [4]. By seeing likes and dislikes, it can be very easy to judge the quality of the product. The quality of the product also be judge by reading the previous analysis.

They said reviews are very important for the new users and the manufacturer company of the relevant product. It helps the company to raise their standards according to the users need and requirements. It can also help the new users to think about the about the brand by reading the reviews of the old users (the people who already used that product) [5]. The number of users are increasing gradually on e-marketing and e-business by the facility of the internet and online shopping. With the increase of the online users who used the products and give their feedbacks. The number of evaluation are also increasing with it. In current time, internet has revolutionized the communication world into a global village [6]. The few companies existed in the past were generating or gathering data (Amazon, Microsoft, Google, and others). In the previous era, the
remaining world used that data. These days, every person is using and create data [7]. The quintillion bytes of data is being produced on daily basis as the previous studies stated that [8]. To handle the huge amount of data is not a piece of cake. This information linked to any data is very useful to know about its past. By using this data information, we can easily sympathize “What has happened in the past” and to predict “What will happen?” [9]. Information has a crucial role in any analysis to give recommendation or making predictions.

Over the last few decades, it is observed that the entertainment platforms of sub-continent had evolved massively [10]. In its enrichment, internet technology plays a vital role. Presently, people have an easy access to watch live matches, anthems of the different leagues (i.e. cricket, football, and hockey) songs, dramas, movies, and seasons and much more on internet using live streaming. Interesting thing is that all that materials are available on live streaming and can be view online, where the spectators give their feedbacks and assess to present their feelings. There exists different websites, blogs and live streaming channels where these languages and other entertainment materials presents with the help of videos that consists of songs, seasons, movies, lectures, documentaries, plays etc. YouTube is one of such platform where any video can easily be accessed [11]. YouTube provides facility for their users, to examine the entire video in detail [12]. By reading these judgements, it can easily be known that what type of content is presented in the video. This is manually feasible only if the number of these analysis are not too high. However, if these considerations are huge in number then we need a proper mechanism to investigate such huge amount of inspections. The massive increase in the assessing views creates a problem to check their polarity as they are Negative, Neutral, or Positive manually. Arises in the reviews need a proper automated mechanism to examine them easily. For this purpose, it is necessary to collect the given information first [13].

II. Related Work

The past studies which were already related to the current piece of study and also fall under the domain of Sentiment Analysis are discussed in details in this chapter. In [14] this paper, they adopted the supervised method to conduct sentiment analysis of reviews on Roman Urdu. However, they also used corpus or dictionary-based techniques to train the machine
learning classifiers on the labeled dataset. Whereas, they perform sentiment analysis on pre-existing dataset of hotel reviews in Roman Urdu. Furthermore, the maximum of the data is written already in Roman Urdu and some of the English reviews are translated into Roman Urdu by using different API’s. While, dataset contains 1600 labeled hotel reviews. They also used dictionary to continue a collection of words to find that whether the probability term of the review is Positive or Negative. Also, they used Hashing Vectorizer to combine Hashing with the pre-processing and Tokenization feature of a Count Vector. From 1600 reviews, they take 1200 for the classifier input and remaining 400 to determine the accuracy of the classifier. Chi-square test is applied on 1200 reviews to check the performance of the classifier by varying the value of “K” training time of the classifier always decreases when “K” decreases. At 1150 reviews behavior of the classifier showing some better results. With HashingVectorizer KNN gives 81%, SVM up to 96%, and Decision Tree gives 77% accuracy, and with CountVectorizer KNN gives 79%, SVM gives 91%, and Decision Tree has 80% accuracy.

In [2] this paper they perform sentiment analysis of Roman Urdu opinions by using different classification algorithms like Naive Bayesian, Decision tree, and KNN. To perform analysis they take 300 reviews on mobile in Roman Urdu that further consist of 150 positive and 150 negative comments. They also used WEKA for the classification techniques and the data file format which they adopted was “ARFF”. They used text dictionary loader command of WEKA in simple “CLI” mode. They train each model on the basis of labeled 150 positive and 150 negative data. Naive Bayesian results on training data gives 95.5% on positive, 99.3% Negative, and overall 97.4% precession. Recall results 99.3% on positive, 95.3% on negative, and ever recall results on training is 97.3%.

Overall results of Naïve Bayesian on training data gives 97.4% precession, 97.3% recall, 97.3 % recall, 97.3% F-measures, and on testing data results of Naïve Bayesian, precession is 97.6%, Recall is 97.3% and F-measures is 97.5% decision tree results on training data 94.7% precession 94.7% recall and f-measures gives 94.7% results while the results on testing data precession gives 92.6%, recall gives 92.5% and F-measures results are 92.5% and the results of
KNN on training data 86.8% precession 86.7 Recall and F-measures gives 86.7%, while on testing data KNN precession 95%, recall all 95% and f-measures also give some Results. The overall accuracy of the Naïve Bayesian on training data is 90% and testing data accuracy is 97.50%. Decision that gives 94.67% accuracy on training and 92.50% on testing KNN gives 86.67% on training and 95% testing accuracy.

In [15] this paper they perform sentiment analysis on the basis of aspect levels. They work on the mobile reviews and get dataset from Amazon web services. The current study extracted reviews from Amazon web services and then extracted the term or aspect they wanted to analyze, later. The important task in aspect level sentiment analysis is data pre-processing they pre-process by using different techniques. They select the set of features to rate on and determine the ratings for selected features. The comments are based on the sentiments where they are summarized according to their rating of features either in positive or in negative manner through the use of SentiWordNet. Whereas, they use data mining association rule to mine the rule and for the segmentation of the sentence, NL process program is used. Finally total numbers of positive and negative comments for each feature were simply counted right after the opinion orientation of the words. Ultimately, they rated the aspect from top to bottom in the level list which already had maximum numbers of positive reviews. At last, rating had been performed according to the nature of aspects that which get minimum positive reviews or which get maximum negative reviews.

In [16] this paper they discussed the sentiment analysis of movie reviews on aspect level classification. However, they discussed it further according to the author’s sentiment analysis is the task to identify the subjectivity in natural language. Moreover, they classified these sentiments at aspect level as well as at document level by exploring a sentiment based scheme. Although, the classification of document level has some linguistic features ranging from adverb + adjective to adverb + adjective + adverb combination. While, in this paper they advised a domain specific heuristic approach for aspect level classification. Also, they extracted the movie reviews from different sites. Whereas, feature extraction can be performed based on SentiWordNet and it would extract the terms having desire POS (Part of Speech).
Finally, it classifies the reviews by using public library in positive, negative and neutral categories and by scoring it likewise. Gram technique is used to indicate the term Aspect 5. While sentiment analysis was conducted, they proposed features which were based on heuristic scheme for the aspect level of movie reviews.

In [17] this survey, the main focus would be on aspect level sentiment analysis with the goal to find and aggregate peoples’ opinions about different entities/products mentioned within the documents. However, during aspect level sentiment analysis we can take a single aspect as a single entity to analyze them. Also, they perform the procedures of identification, classification, and aggregation of sentiment analysis on aspect level. As overall review would generally refer to the entity during this analysis, the aspect detection is the major task or important part in aspect level sentiment analysis. The current study discussed different ways of aspect detection such as, frequency based and syntax based aspect detections. These aspect detections were mainly focused on frequencies of the aspect and by performing syntax based methods they would find the means of the syntactical relaxation in which aspects are already present. Furthermore, this survey would talk about the classification methods which were supervised as well as unsupervised. Supervised classification was done on the basis of labeled data to conduct both training and testing procedures. Unsupervised classification was required to operate labeled data only for training sessions of the algorithm. In addition to it, while testing it can classify the unlabeled data. Hence, state of the art based on aspect level sentiment analysis was presented in this survey.

In [18] survey was carried out on dimensional reduction for SA by using different pre-processing techniques. In this case, they adopted supervised methods to analyze the results of reviews. As we know that pre-processing is a way to clean the data and results which would be increased after pre-processing it as compare to the unprocessed one. In this way, they applied classifiers on both kind of data, either processed or unprocessed. Moreover, they used that dataset which was available on Kaggle called Bag of words, “Bag of Popcorns”. It contains 25000 labeled reviews, out of which 12500 are positive and 22500 are negative reviews. They use random forest classifier which is imported from Sci-Kit learn for tracing the
classifier 22500 reviews they choose and remain 2500 for testing. Total 256 experiments were conducted and first experiment was contended with the combination on unprocessed data. Moreover, the accuracy rate without pre-processing was found to be 83.417%, while on the other hand, they applied pre-processing techniques one by one to achieve the authentic results. Furthermore, first experiments after applying pre-process techniques in which they apply HTML tags techniques the accuracy will increase up to 84.51%. In the second experiment, they remove slangs and apply lemmatization on the accuracy rate to raise it up to 85.74%. In third experiment, they apply three pre-processing techniques including slangs handling, stop words removed, and lemmatization combined with accuracy they achieved the maximum accuracy from all the experiments they performed is 86%. In fourth experiments, slangs handling, stop words removed, stemming, and lemmatization had the accuracy 84.9%. In fifth combination of experiments, the maximum accuracy is 84.59%, in 6th experiment 84.468%, in 7th experiment 84.34%, and in the last 8th combination of experiments, eight pre-processing techniques had one experiment with 84.831% accuracy.

In [3] this paper, the current study is based on the discourse on which sentiment analysis is being performed through Natural language processing on Roman Urdu dataset. Furthermore, they gather datasets from different blogs, sites, news, and they also gather approximately 21000 reviews from newspapers. Discourse is that form of natural language which can be studied as ‘language of text’ that deals with the understanding of how and why the writings of some users have more impact on hearers than on readers. They take 80% data form training and remaining 20% is test data. They are more focused to generate results from the collected data corpus. As this is a challenging collection of scientifically large data corpus. They also used dataset comprising of statement written in Roman Urdu and used F1 score calculation for evaluating the accuracy of the results. They also applied precession and results to compare the results they get. They adopted POS (Part of Speech) tag for each entity with Google APIs.

Discourse based POS tagging results gave different accuracy rates based on different types of datasets. These reviews were from “Bio Social Workers” which attained
78% of success rate, bio geographic has 90%, and reviews taken form Shasha shows brilliant results around 100% success rate and so on. Along with the precession 84%, and recall 90.6%, they gave and bio social workers data, bio graphics has 0.92% precession, and 97% recall, and at Shasha overall the 100% correct results in precession, Recall and with F1-Score all dataset shows accuracy of around 80% on an average.

In [19] survey is carried out on the detection of reviews from sentiments. As sentiment detection or sentiment analysis of reviews in texts has a booming interest, so a large number of reviews can be found on different blogs. To analyze such reviews, there is a need to perform SA on these reviews. The SA analysis can be performed in different ways, that is, document, sentence, and aspect level sentiment analysis. The sentiment analysis can be done in two ways including binary type of data, such as positive and negative reviews. It can also be done through polynomial type of data, that is, positive, negative and neutral.

In [20] solution is proposed of different forms, by performing multiple functions for Roman Urdu dataset which was collected by conducting a survey in local universities of Pakistan. While conducting the survey, total participants were 116 in number, (58 male and 52 female). Average age of both male and female was 21.01 years in which 103 (88.8%) participants were taken from undergraduates 10 (8.6%) were taken from graduates. While 3 (2.6%) of them were PhD students. The messages they collected were in the form of text messages and corpus had total 4,46,483 words. During their survey, they observe those people who mostly prefer to write their messages in Roman Urdu. According to the current study, they analyzed that the female data was written in less romantic words than males. Moreover, the undergraduates have used more intimate words than the graduates. 73 users, used 20 or less intimate words which are called low romantic participants. 31 such users that use 21 to 80 intimate words they are classify as medium Romantic Peoples. The 41 users are those who used 81 or more intimate words they are classify as high romantic participants. The current research was carried out to understand the population ways of messaging and classifying the users who are adopted in low, medium, or high romantic manner to communicate with others.
In [21] this paper, they carry a survey on supervised and unsupervised classification of documents. This classification of documents aim to poll them into further classes. For example, they poll two documents one as positive and second as negative. There are three basic methods to classify these document such as rule based, supervised, and unsupervised document classification. In rule based classification of documents, the defined rules are query phrases, and in supervised classification, the documents are classified on the basis of supervised learning. Actually, they first train the algorithm by using the training data which is known as ‘algorithm learning’. After learning the algorithm, they properly classify the testing data. On the other hand, supervised classification has two type of content request based classifications. It can simply cluster the data according to its pattern or structure. Later, unsupervised algorithms were discussed under supervised centered based approach, support vector machine, Naive Bayes classifier, and decision tree. In centered based techniques, each documents "D" represented Documents Vector VD and centered vector of each class, and Euclidean distance between VD, and centered Vector of Class is calculated. Hence, documents having minimum distance from centered based approach assign a particular class.

SVM analyzes the data and classifies them after recognizing the pattern. Naive Bayes classifies the document by the calculation of posterior probability value and classify the documents into that particular class which has the highest form of frequency. Decision tree uses tree based algorithm to classify the documents. In unsupervised classification they discussed partitioned clustering. In partition clustering, un-nested partition was used and K clusters are defined and partitioned, initially. Remain (P) is constructed, then redefined clustering solution is provided by moving the documents form one cluster to another, iteratively. In K-mean clustering, K clusters are defined and in this situation, each documents move to that cluster which is nearest to its center. Hence, hierarchical clustering techniques make that cluster which is moving from top to bottom and then again to upward direction where documents are divided into further clusters.

**III. Research Methodology**

As the third phase is about dataset collection and generation. Furthermore, in the fourth phase the pre-processing techniques are
discussed in details which are used to make data free from noise and make it ready for analysis. In addition to it, the fifth phase is about the post-processing where the machine learning techniques are used to classify the Roman Urdu reviews. In the sixth phase, the results are discussed in details and finally the current study is concluded. While the future work is listed in this case that would be further extended in next years.

Although, data plays a vital role in the current research analysis, it is considered to be the backbone of any analysis \[22\]. Whereas, collecting data/information is not a trivial component. Also, it requires efforts to collect massive amount of data. Moreover, there exist various IR techniques that deals with the huge amount of data collected from different resources \[23\].

| SR# | Anthems URLs | Release Date | No # Views | Total comments |
|-----|--------------|--------------|------------|----------------|
| 1   | https://www.youtube.com/watch?v= YwFlKiWSCTk | Sep 14, 2016 | 19,825 | 319 |
| 2   | https://www.youtube.com/watch?v= NFqzetIVKrQ | Jan 30, 2017 | 18,821,877 | 12408 |
| 3   | https://www.youtube.com/watch?v= W3eKzfd1p24 | Jan 29, 2018 | 8,204,165 | 10925 |
| 4   | https://www.youtube.com/watch?v= Xr_1ZbEk8Y8 | Jan 18, 2019 | 15827131 | 23572 |
| 5   | https://www.youtube.com/watch?v= 7nOR4XEL47M | Jan 28, 2020 | 7,687,963 | 18,201 |
| 6   | https://www.youtube.com/watch?v= KBdjwrywyKo | Feb 6, 2021 | 18642378 | 72503 |

In this particular research, we are going to analyze Roman Urdu reviews on PSL anthems. Initially, it’s compulsory to...
analyze the collection of reviews and also the reviews collected through scraper from YouTube [24] on first to sixth PSL anthems. The collected review data is saved in CSV file format right after scraping them from the YouTube comment section.

First six official anthems of the PSL were selected to collect the reviews and feedback of people. The URLs of the anthems with their release date, the total number of views on each anthem, and the total number of reviews (raw data, the reviews with noisy data) that are collected from the video are mentioned in the Table I.

A. Data Labeling

Data annotation is a static part and it has a trivial component of Sentiment Analysis. Data annotation is a process to categorize the instance/review/comment into Positive, Neutral, or Negative manner [25]. Further, in the data annotation process, each review/comment assign a class according to the subjectivity. The current study applies Sentiment Analysis on Roman Urdu by using supervised Machine Learning Algorithms. In this case, the data must be in label form for supervised learning. Whereas, the labelling of review is also known as data annotation [26]. Previous studies showed and highlighted that there were different ways to manually annotate the reviews. Also, a class was assigned to those reviews according to its subjectivity. For this particular piece of study, the manual annotation has been performed. In manual annotation, each review was labelled by reading them one by one and also they were assigned with appropriate class according to its behavior in Positive, Neutral, and Negative manner [27]. As, the manual annotation is performed by involving annotation. Also, the second phase is the training phase and the third phase is conflict resolving and computation of inter-annotator agreement. In the annotation process, the three persons are involved, as they are all graduates and the native speakers of Urdu language. Furthermore, they are all well aware with the task of sentiment analysis and the way of annotation of the reviews [28]. The annotation guidelines are as follows:

B. Annotation Guidelines

However, the guidelines have been prepared in this case to annotate the anthems and
reviews/comments of the viewers on YouTube. In addition to it, the guidelines are provided for each class demonstrating them separately. Guidelines for positive and negative classes are mentioned as follows:

Positive class has been assigned to those reviews which belong to some positive behavior of people [29]. If a review shows neutral behavior as well as positive, the class assigned to them is also positive [30]. The presence of a positive word in a sentence is also marked as positive [31]. A review having illocutionary speech acts like wow, congrats, and smash are also classified as positive [32], that is, a review “tarana acha h” in which reviewer showed his positive gratitude towards the anthem which has been highlighted positively by a word “acha” that clears the polarity of the review. Another review “kamal ki awaz h” in which reviewer is being positive towards the voice of the singer which is shined by the word favorite.

The negative class has been assigned to those reviews which have some negative terms of sentiments. If a review is having an abusive information, it would be classified as Negative. If a reviewer shows an un-softened behavior towards sentiment, it is also classified as negative [33]. Negation in such a review which makes more negative in sense, that is, a review “bakwas e h sb” in which the word “bakwas” clears the polarity of the review is negative. Another review “sb fazool h” in which the word “fazool” clears the polarity of the review towards negative connotation. Whereas, after making the guidelines, the annotators were provided with the training that how to annotate the data through examples by following instructions. Manual annotation is performed by the three annotators labelled as A, B, and C. After these training phases, the pre-processed data are given to annotators to annotate each of the review. However, the inter-annotator value is calculated by using “Kappa Statistics”. In this case, all the conflicts were resolved between the annotators and finally all the reviews were annotated manually by reading all the reviews and comments one by one. In practice, initially 200 reviews were given to the annotators which are labelled as “A and B” to annotate by following its primary version of annotation of guidelines. After annotating these 200 reviews, maximum conflicts were resolved in the meeting of all three annotators.
Also, the guidelines were updated and then 100 reviews were given to the annotator “C” and after the annotation, the inter-annotator agreement was calculated that was 87.00% by using “Kappa Statistics”. Ultimately, this was a good and acceptable score. Finally assigned polarity to each review as Positive and Negative. A sample annotated data is shown in Table II.

Table II
Sample Label Dataset

| Comments                                           | Class   |
|----------------------------------------------------|---------|
| Tera apna level ha Teri jaga koi ni ly skta       | Positive|
| Bohttt kamaaaal song                               | Positive|
| Is anthem ka to level hi alag hai                  | Positive|
| Yeh hota hai song                                  | Positive|
| Groove Mera Mera                                   | positive |
| Tera apna level ha Teri jaga koi ni ly skta       | positive |
| yrrr it still sounds so goooooood                  | positive |
| After aagy dykh it is still our favorite           | positive |
| Sorry Standard bohat ghatya h                      | negative |
| Bhoot ghatya                                       | negative |
| Atiff Aslam ka is se better ha                     | negative |

The final corpus consisted of 7000 Bi-nomial reviews which were manually annotated right after involving three annotators. Hence, the final annotation which was about gold standard corpus, belongs to the two classes which are positive and negative. The traits of the final corpus are mentioned in Table III.

Table III
Final Corpus Traits

| Characteristics                          | Traits        |
|------------------------------------------|---------------|
| Number of Reviews Scraped                | 137928        |
| Filtered Roman Urdu Reviews              | 7000          |
| Reviews of Positive Class                | 3500          |
| Reviews of Negative Class                | 3500          |
| Total Number of Characters in Dataset    | 330947        |
| Maximum Length of a review in character  | 385           |
| Minimum Length of a review in character  | 03            |
C. Classification

The classification of sentiments can be done in the form of two different types. The poly-nominal type of classification and Bi-nominal type of classification.

In this particular research, the binominal type of sentiment analysis is done by using different supervised Machine Learning Algorithm based on Roman Urdu-Hindi reviews. Gradient Boost Tree (DT), Naïve Bayes (NB), K-Nearest Neighbors (KNN), and Recurrent Neural Network (RNN) algorithms are applied to perform Sentiment Analysis on label data. The model which is designed to classify the reviews is shown in Figure 1.

III. Results and Discussion

In this research, we performed Sentiment Analysis of Roman Urdu reviews regarding PSL anthems posted by their viewers on YouTube in the comment section. These comments are collected from YouTube with help of scraper. To perform Sentiment Analysis we have chosen supervised Machine Learning Algorithm. To evaluate and determine the performance of the classifiers, we opt four different techniques (Accuracy, Precession, Recall, and F-Measure).

Moreover, the collected dataset is scraped from the YouTube channel of the PSL anthems. The first six anthems of the PSL have been selected to perform Roman Urdu Sentiment Analysis. These reviews belong to the two classes of people who comment either in positive or in negative way. After the data is collected and pre-processed, it further undergoes the process of manual annotation. After making data easy for the analysis. The five different machine learning techniques are applied to classify the reviews and to perform sentiment analysis on Roman Urdu reviews.

In the current study, supervised ML algorithm are implemented on Roman Urdu-Hindi reviews for the Bi-nominal (Positive and Negative)
type of reviews. To classify these reviews supervised, Machine Learning techniques are applied. Two of them from basic Machine Learning algorithms (Naïve Bayes and Decision Tree—Gradient Boost Tree) are implemented and one is the regression algorithm SVM, which is used to classify the reviews. One from the clustering in KNN and one classifier is implemented from Neural Net is RNN.

A simple statistical-based probabilistic classifier which represents the supervised learning method is Naïve Bayes. It can simply be based on the “Bayes” theorem. Naïve Bayesian classifier can be trained very fast and efficiently in supervised learning.

Table IV

Confusion Matrix of Naïve Bayes

|       | Positive | Negative |
|-------|----------|----------|
| Pred. |          |          |
| Positive | 330      | 218      |
| Pred.   | 0        | 152      |
| Negative|          |          |

The model was predicted by using the Naïve Bayes and classifying 330 reviews as a true positive, 218 reviews as false positive, and 152 reviews as true negative. Furthermore, the matrix is shown in below Table 4. Naïve Bayesian correctly predicted 68.86% reviews correctly, with the recall of 100%, precision is 60.22%, and F-score is 75.17%. Details are shown in Table V.

Table V

Results of Naïve Bayes

|        | Accuracy | Recall | Precision | F     |
|--------|----------|--------|-----------|-------|
|        | 68.86    | 100    | 60.22     | 75.17 |

1) A Decision Tree performs a hierarchical based classification of the document. In Decision Tree, a document is classified by using the flow-chart structure. The decision is a method which selects one choice out of the two. From the decision tree family, the Gradient Boost Tree is implemented. The gradient predicted 245 reviews as true positive, 167 as false positive, 85 reviews as false negative, and 203 reviews predicted as true negative. Details are shown in Table VI.

Table VI

Confusion Matrix of Gradient Boost Tree

|       | Positive | Negative |
|-------|----------|----------|
| Pred. |          |          |
| Positive | 245      | 167      |
| Pred.   | 85       | 203      |
| Negative|          |          |
2) Gradient Boost Tree accuracy is 64.00%, with the precision of 59.47%, recall is 74.24%, and the value of F-score is 66.04%. Details are shown in Table VII.

Table VII
Results of Gradient Boost Tree

| Accuracy | Recall | Precision | F   |
|----------|--------|-----------|-----|
| 64.00    | 74.24  | 59.47     | 66.04 |

3) Logistic Regression is used for the binomial type of classification. It used statistical based technique to analyze the data. It can simply predict binary outcomes, that is, positive or negative and yes or no category. The Logistic Regression Predicted 328 reviews as true positive, 154 as true negative, 216 as false positive, and 2 reviews as false negative. Details can also be seen in Table VIII.

Table VIII
Confusion Matrix of Logistic Regression

|         | Positive | Negative |
|---------|----------|----------|
| Pred. Positive | 328      | 216      |
| Pred. Negative  | 2        | 154      |

4) The Logistic Regression performs 68.86% results in term of accuracy, 99.39% recall, 60.29% precision, and F-score is 75.06%. Results are also being seen in Table IX.

Table IX
Results of Logistic Regression

| Accuracy | Recall | Precision | F   |
|----------|--------|-----------|-----|
| 68.86    | 99.39  | 60.29     | 75.06 |

5) The K-nearest-neighbor classifier was first introduced in the mid-1950s and did not gain popularity till the 1960s. However, it relies on instance-based learning. KNN algorithm assumes that all the instances are corresponding to the points in the n-dimensional space. KNN predicted 328 reviews as true positive, 328 as false positive, 2 reviews predicted as false negative, and 42 reviews as true negative. It can also be seen in Table X.

Table X
Confusion Matrix of KNN

|         | Positive | Negative |
|---------|----------|----------|
| Pred. Positive | 328      | 328      |
| Pred. Negative  | 2        | 42       |

6) Moreover, KNN performs 52.86% in term of accuracy, F-score is 66.53%, precision, and recall value is 99.39%, and
50.00% respectively. Results can also be seen in Table XI.

Table XI
Results of KNN

| Accuracy | Recall | Precision | F  |
|---------|--------|-----------|----|
| 52.86   | 99.39  | 50        | 66.53 |

7) However, deep learning is simply a feed-forward neural network and it can also be known as hierarchal learning or deep structured learning which is further based on the multilayer feed-forward neural networks. Artificial neural network is implemented to perform deep learning in rapid miner. Hence, the ANN predicted 219 reviews as true positive, 128 as false positive, 111 reviews as false negative, and 241 reviews predicted as true negative. The confusion matrix can be seen in Table 12.

Table XII
Confusion Matrix of ANN

|            | Positive | Negative |
|------------|----------|----------|
| Pred. Pos. | 219      | 128      |
| Pred. Neg. | 111      | 242      |

8) The ANN predicted 65.86% reviews correctly with the value of F-score 64.69%. Where the value of recall is 66.36% and the precision is 63.11%. The results can also be seen in Table XIII.

Table XIII
Results of ANN

| Accuracy | Recall | Precision | F  |
|----------|--------|-----------|----|
| 65.86    | 66.36  | 63.11     | 64.69 |

9) In addition to it, the model is developed by splitting the dataset into 90% for the training and rest of the 10% is used to test the model. Five different machine learning algorithms are implemented including—Naïve Bayes, Gradient Boost Tree, Linear Regression, K-Nearest Neighbors, and Artificial Neural Network. Testing results are shown in Table XIV.

10) At the testing of the model, the NB, and KNN outperforms the other classifiers with the accuracy of 68.86%. The value of F-score is 75.17%, and 75.06% respectively. The second closest accuracy achieved by the ANN, which is 65.86%, and the F-score is 64.69%. Anyhow, rest of the results comparison is shown in Figure 2.
|       | Accuracy | Recall | Precision | F    |
|-------|----------|--------|-----------|------|
| KNN   | 52.86    | 99.39  | 50        | 66.53|
| ANN   | 65.86    | 66.36  | 63.11     | 64.69|

In addition to it, the accuracy of Naive Bayes and Logistic Regression is same and the highest in term of accuracy as compared with all other classifiers. As the results of ANN at testing data and validation of results are very closer to each other. Also, ANN achieved the highest accuracy at the validation of the model. Hence, this particular study recommends ANN for Sentiment Analysis of Roman Urdu Reviews.

### IV. Conclusion

Sentiment analysis of human language is an emerging field that is growing exponentially in research domain. Sentiment Analysis is an effective way to analyze the opinions about the text which are posted by the people on different blogs and on web pages. This research was conducted through Sentiment Analysis of Roman Urdu reviews. To perform this research, the first six anthems of PSL were selected for textual analysis. Reviews that were posted by their viewers on YouTube in the comment section were scraped with the help of scraper. Scrapped reviews were in multi—languages, Roman Urdu reviews were separated from other languages by applying different filters. After filtering the Roman Urdu reviews, pre-processing was performed to collect
and organize the data. These reviews underwent the process of manual annotation after pre-processing the data, even if reviews were positive or negative in nature. After the process of annotation was completed, the dataset was provided with explanatory notes or comments fully. In this way, speech data was made free from all type of noises. In the current study, the machine learning classification techniques are implemented to classify the reviews. The five different machine learning algorithms, that is, NB, GB, SVM, KNN, and ANN are implemented to perform Roman Urdu Sentiment Analysis. The Naïve Bayes and Logistic Regression achieved 68.86% accuracy at testing of the model. ANN classify the 68.86% reviews correctly with the value of 64.69% F-score and at the validation of results ANN achieved 69.71% accuracy and 76.64% F-score. The testing and validation accuracy of ANN is approximately the same and ANN outperforms the other classifiers at the validation of the model. So, this study recommends ANN for the Sentiment Analysis of Roman Urdu Reviews.

References

[1] M. A. Qureshi et al., “A Novel auto-annotation technique for aspect level sentiment analysis,” Comput. Mate. Continua, vol. 70, no. 3, pp. 4987–5004, 2022.

[2] M. Bilal, H. Israr, M. Shahid, and A. Khan, “Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques,” J. King Saud Uni. Comput. Inf. Sci., vol. 28, no. 3, pp. 330–344, Jul. 2016, doi: https://doi.org/10.1016/j.jksuci.2015.11.003

[3] Z. Sharf, D. Saif, and U. Rahman, “Performing natural language processing on roman urdu datasets,” IJCSNS Int.J.Comput. Sci. Net. Sec., vol. 18, no. 1, pp. 141–148, 2018.

[4] M. A. C. Jondar, “Rich Youtuber, Poor Youtuber: Implementasi business intelligence dalam meningkatkan pendapatan channel youtube YE,”Undergraduate thesis, Surabaya, 2020. [Online]. Available: http://digilib.ubaya.ac.id/pustaka.php/258409

[5] R. J. Varey, “Internal marketing: a review and some interdisciplinary research challenges,” Int. J. Service
[6] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, “Discriminative feature spamming technique for roman Urdu sentiment analysis,” IEEE Access, vol. 7, pp. 4799-48002, 2019, doi: https://doi.org/10.1109/ACCESS.2019.2908420

[7] J. Hendler, “Data integration for heterogenous datasets,” Big Data, vol. 2, no. 4, pp. 205–215, Dec. 2014, doi: https://doi.org/10.1089/big.2014.0068

[8] G. Bello-Orgaz, J. J. Jung, and D. Camacho, “Social big data: Recent achievements and new challenges,” Information Fusion, vol. 28, pp. 45–59, Mar. 2016. https://doi.org/10.1016/j.inffus.2015.08.005

[9] P. Hewson, “Reconsidering conceptual change: Issues in theory and practice,” Sci. Edu., vol. 87, no. 6, pp. 913–916, Nov. 2003, doi: https://doi.org/10.1002/sce.10121

[10] M. Asif, M. A. Qureshi, A. Abid, and A. Kamal, “A Dataset for the Sentiment Analysis of Indo-Pak Music Industry,” in 2019 Int. Conf. Innov. Comput. (ICIC), 2019, pp. 1–6, doi: https://doi.org/10.1109/ICIC48496.2019.8966720

[11] S. Zhang, T. Aktas and J. Luo, "Mi YouTube es Su YouTube? Analyzing the Cultures using YouTube Thumbnails of Popular Videos," 2021 IEEE Int. Conf. Big Data, 2021, pp. 4999-5006, doi: https://doi.org/10.1109/BigData52589.2021.9672037

[12] R. Raby, C. Caron, S. Théwissen-LeBlanc, J. Prioletta, and C. Mitchell, “Vlogging on YouTube: the online, political engagement of young Canadians advocating for social change,” J. Youth. Stud., vol. 21, no. 4, pp. 495–512, 2018, doi: https://doi.org/10.1080/13676172017.1394995

[13] A. Madden, I. Ruthven, and D. McMenemy, “A classification scheme for content analyses of YouTube video comments,” J. Document., vol. 69, no. 5, pp. 693–714, Sep. 2013, doi:
[14] O. Beijbom et al., “Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation,” *PLoS ONE*, vol. 10, no. 7, pp. 1–22, July 2015, doi: https://doi.org/10.1371/journal.pone.0130312

[15] E. Gur-esh, N. Fink, D. Friesenhahn, N. Ramkumar, and M. J. FitzGerald, “Techniques for managing persistent document collections.” US patent 8,965,926 B2, Feb. 2015.

[16] Z. Mahmood et al., “Deep sentiments in Roman Urdu text using recurrent convolutional neural network model,” *Inf. Proc. Manag.*, vol. 57, no. 4, pp. 102233–102246, Jul. 2020, doi: https://doi.org/10.1016/j.ipm.2020.102233

[17] L. Nemes and A. Kiss, “Social media sentiment analysis based on,” *J. Inf. Telecommun.*, vol. 5, no. 1, pp. 1–15, Jul. 2021, doi: https://doi.org/10.1080/24751839.2020.1790793

[18] R. G. Curty and P. Zhang, “Social commerce: Looking back and forward,” in *Proc. ASIST Ann.Meeting*, vol. 48, no. 1, pp. 1–10, Jan. 2011, doi: https://doi.org/10.1002/meet.2011.14504801096

[19] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, “Comparing automated text classification methods,” *Int. J.Res. Market.*, vol. 36, no. 1, pp. 20–38, Mar. 2019. https://doi.org/10.1016/j.jiresmar.2018.09.009

[20] C. Zhang, D. Zeng, J. Li, F. Y. Wang, and W. Zuo, “Sentiment analysis of chinese documents: From sentence to document level,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 12, pp. 2474–2487, Dec. 2009, doi: https://doi.org/10.1002/asi.21206

[21] A. Shoukry and A. Rafea, “Sentence-level Arabic sentiment analysis,” in *Proc. 2012 Int. Conf. Collab. Technol. Sys.*, CTS 2012, May 2012, pp. 546–550, doi: https://doi.org/10.1109/CTS.2012.6261103

[22] Y. Zhang, Z. Zhang, D. Miao, and J. Wang, “Three-way
enhanced convolutional neural networks for sentence-level sentiment classification,” *Inf. Sci.*, vol. 477, pp. 55–64, Mar. 2019, doi: [https://doi.org/10.1016/j.ins.2018.10.030](https://doi.org/10.1016/j.ins.2018.10.030)

[23] T. Chen, R. Xu, Y. He, and X. Wang, “Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN,” *Exp. Sys. App.*, vol. 72, pp. 221-230, Apr. 2017, doi: [https://doi.org/10.1016/j.eswa.2016.10.065](https://doi.org/10.1016/j.eswa.2016.10.065)

[24] M. del P. Salas-Zárate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M. A. Rodríguez-Garcia, and R. Valencia-Garcia, “Sentiment analysis on tweets about diabetes: An aspect-level approach,” *Comput. Math. Methods Med.*, vol. 2017, pp. 1–9, 2017, doi: [https://doi.org/10.1155/2017/5140631](https://doi.org/10.1155/2017/5140631)

[25] F. Chen, Z. Yuan, and Y. Huang, “Multi-source data fusion for aspect-level sentiment classification,” *Knowledge-Based Sys.*, vol. 187, pp. 104831, Jan. 2020, doi: [https://doi.org/10.1016/j.knosys.2019.07.002](https://doi.org/10.1016/j.knosys.2019.07.002)

[26] M. Yang, W. Yin, Q. Qu, W. Tu, Y. Shen, and X. Chen, “Neural attentive network for cross-domain aspect-level sentiment classification,” *IEEE Trans. Affec. Comput.*, vol. 187, Art. no. 104831, 2019, doi: [https://doi.org/10.1109/taffc.2019.2897093](https://doi.org/10.1109/taffc.2019.2897093)

[27] M. Syahrul and M. Dwi, “Aspect-based sentiment analysis to review products using Naïve Bayes,” *AIP Conf. Proc.*, 2017, doi: [https://doi.org/10.1063/1.4994463](https://doi.org/10.1063/1.4994463)

[28] H. H. Do, P. W. C. Prasad, A. Maag, and A. Alsadoon, “Deep learning for aspect-based sentiment analysis: A comparative review,” *Expert Syst. Appl.*, vol. 118, pp. 272–299, Mar. 2019, doi: [https://doi.org/10.1016/j.eswa.2018.10.003](https://doi.org/10.1016/j.eswa.2018.10.003)

[29] L. E. Ribeiro, D. W. Tokikawa, J. L. Rebelatto, and G. Brante, “Comparison between LoRa and NB-IoT coverage in urban and rural Southern Brazil regions,” *Ann. Telecommun.*, vol. 75, no. 11,
Sentiment Analysis of …

pp. 755–766, Jul. 2020, doi: https://doi.org/10.1007/s12243-020-00774-3

[30] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, “Sentiment analysis for a resource poor language-Roman Urdu,” in ACM Trans. Asian Low-Resource Lang. Inf. Proc., vol. 19, no. 1, pp. 1-15, Jan. 2020, doi: https://doi.org/10.1145/3329709

[31] R. Habiba, D. M. Awais, and D. M. Shoaib, “A Technique to calculate national happiness index by analyzing Roman Urdu messages posted on social media,” in ACM Trans. Asian Low-Resource Lang. Inf. Proc., vol. 19, no. 6, Nov. 2020, doi: https://doi.org/10.1145/3400712

[32] M. Kubat, An introduction to machine learning, vol. 35. Cham: Springer International Publishing, 2017, doi: https://doi.org/10.1007/978-3-319-63913-0

[33] S. Kramer and C. Helma, Machine learning and data mining, vol. 42, Woodhead Publishing Limited, 2005, doi: https://doi.org/10.1533/9780857099440