ISSN (Print): 0974-6846
Indian Journal of Science and Technology, Vol 12(35), DOI: 10.17485/ijst/2019/v12i35/146571, September2019 ISSN (Online): 0974-5645

Roman-Urdu News Headline Classification with IR Models using Machine Learning Algorithms

Syed Muhammad Hassan[1*], Fayyaz Ali[2], Shaukat Wasi[3], Samreen Javeed[1], Imtiaz Hussain[1] and Syeda Nazia Ashraf[1]

[1]Department of computer science, Sindh Madressatul Islam University, Karachi, Pakistan
M.hassan@smiu.edu.pk, samreen.javeed@smiu.edu.pk, Imtiaz@smiu.edu.pk, snazia@smiu.edu.pk
[2]Department of computer science, Sir Syed University of Engineering and Technology, Karachi , Pakistan; Fayyaz54@gmail.com
[3]Department of computer science, Mohammad Ali Jinnah University, Karachi, Pakistan; Shaukat.wasi@jinnah.edu

Abstract

Objectives: Roman-Urdu consider as a non-standard language used frequently on the Internet. To classify text from article tagging on Roman-Urdu is such difficult task because of many irregularities in spellings, for example, the word khubsurat (beautiful) in Roman-Urdu has multiple spellings. It can also be written as khoobsurat, khubsoorat, and khobsorat. **Methods/Statistical Analysis:** In this study, we scrap Roman-Urdu language news headline from various online newspapers. Our corpus contains 12319 news headlines which contain seven categories i.e. Accident, Sports, Weather, Arrest, Conference, Operation and Violence. We also use different preprocessing approaches like Roman-Urdu Stop words and apply IR models i.e. TF-IDF and Count Vector for feature extraction before applying classifier algorithms. **Findings**: We also compare results between different Machine Learning algorithm such as RF, LSVC, MNB, LR, RC, PAC, Perceptron, NC, SGDC and NC. Our model predicts best result to identify desire class on SGD classifier which gives 93.50% accuracy. **Application/ Improvements:** It is recommended that SGD Classifiers should be used in roman-Urdu news headline text classification.

Keywords: Linear SVC, Multinomial Naïve Bays (MNB), Ridge Classifier (RC), Random Forest, Roman-Urdu, Supervised Machine Learning, Stochastic Gradient Descent (SGD), Text Classification, Tf-Idf

1. Introduction

Large amount of data with its all variations on internet is available nowadays; most interestingly languages are no more barriers to identify information. Many people are interested to get their knowledge in the form of their native speaking or written languages. Roman-Urdu is one of the most popular and increasing demanding language nowadays with blend of English and Urdu[1]. To analyze text with its category is most common and useful technique that cover all major field of Natural Language Processing for example sentiment analysis, opinion mining, reviews, tweets, blogs, spam detection and something whose sentiment is to be evaluated. Two major methods use for classification textual data are corpus-based[2] in which pre-define dictionary uses that contains collection of words and lexicon-based finds the polarity of every word or phrase in a text document. Considering the popularity of Roman-Urdu, we make a model to classify news headline based on Roman-Urdu language on our own captured corpus. To analyze dataset, we use seven different categories as shown in Figure 1 named as Accident, Sports, Weather, Arrest, Conference, Operation and Violence news take as an input and passes into ten different machine learning algorithms to predict desire class. Furthermore, TF-IDF features vectors visualized through t-SNE plotting Graph as represent in Figure 2. Our method contains different primary

*Author for correspondence

processes: stop words removal, feature vector, which use to predict class for sentences by applying the machine learning algorithms. Most of researchers previously work on Roman-Urdu in the context of sentiment analysis and opinion mining with limited number of Supervised Machine Learning Algorithms such as Naïve Bays (NB), Logistic Regression with Stochastic Gradient Descent (LRSGD) and Support Vector Machine (SVM). We execute 10 Machine Learning Algorithms on seven

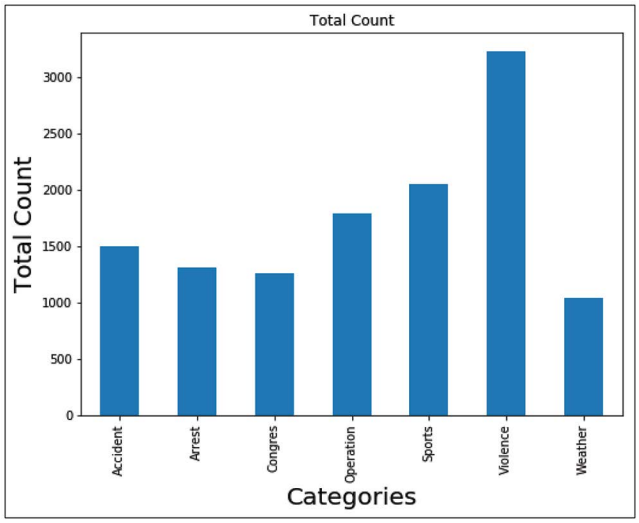

Figure 1. Dataset size category-wise.

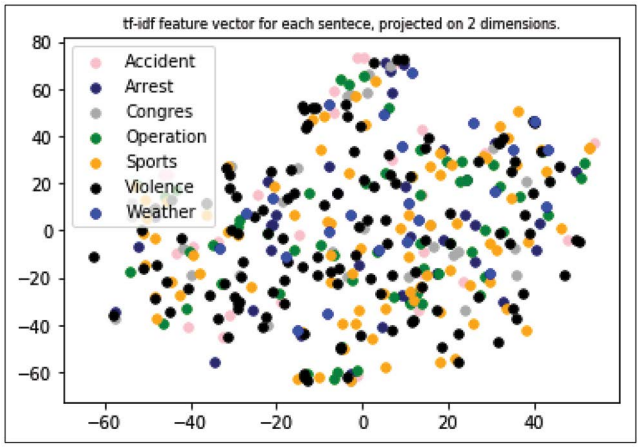

Figure 2. TF-IDF feature category-wise.

different categories.

This study has multiple sections, Section 2 describe related work of different researchers on classifying text, topic modeling, sentiment analysis and text mining. Section 3 belong to methodology which describe the whole procedure to collect data using web crawler, preprocess data, feature extraction using TF-IDF and vector model. Section 4 describes results that predict our class on any given sentence and last Section 5 contain conclusion.

Waikato Environment for Knowledge Analysis (WEKA) tool to analyze different opinions written in Roman-Urdu[1] and English from a blog for text classification. Balance data use for both classes, negative and positive each contains 150 opinions that are documented in text files which consider as a training dataset. Three machine learning models is use for training and testing, in which Naïve Bayesian performance is far better than Decision Tree and K-Nearest Neighbor (KNN) after analyzed accuracy, precision, recall and F-measure. Ayesha et al. Online remarks, feedback or any type of opinion from public on specific domain is nowadays very common. To analyze human cognitive behavior and to understand user opinions in order to understand latest trend is refer to as sentiment analysis[2]. This study is to analyze user hotel reviews by applying different classifiers and feature selection and representation with the conversion of English dataset into Roman-Urdu corpus. The proposed methodology is to analyze customer feedback that assists organization to improve their product, services, and marketing strategies. Computational Linguistic or NLP are diversifying field due to ambiguity in speech and language processing[3]. To analyze data Machine learning (supervised or unsupervised) and statistical techniques are powerful tools for various NLP tasks. The aim of this study is to classify different categories using different machine models i.e. Hidden Markov Models (HMM), Conditional Random Field (CRF), Maximum Entropy models (MaxEnt), SVM and NB on ambiguity in speech and language processing which identify best techniques to apply on linguistic knowledge. Compare standard[3] text and lexical resources complexities during processing text on resource poorer language Urdu[4] and resource rich languages for carrying various NLP tasks in any languages of the globe. This study compares rule based and statistical methods performance on developing large annotated datasets using statistical learning for Urdu Language Processing. As a result, statistical methods perform better due to low amount of large annotated datasets that require testing performance on Urdu Language and Other Languages Processing. Analyze Urdu language text by capturing data from different blogs and annotated with the help of human

annotators. After annotated data it passes through well know machine learning algorithm i.e. SVM, Decision tree[5] and KNN to find their performance in term of accuracy, precision, recall and f-measure. In[6] classifying text with seven categories (Business, Entertainment, Culture, Health, Sports, and Weird) on Urdu corpus contains 21769 news documents. For predict classes different machine learning algorithm apply on 93400 features extracted from dataset which gives 94% precision and recall using classify class. Apply Deep learning model called Long-Short Time Memory (LSTM) on Roman-Urdu dataset to analyze sentiments with the comparison of Machine Learning methods. Result shows that deep neural networks-based model is perform better on sequential data models without applying preprocessing techniques as compared to Machine leaning approaches. Also suggest that word embedding with LSTM is successful approach to perform Sentiment Analysis[7]. Discuss major issues related to text such as handling large number of features, unstructured text and textual content with solution by applying appropriate semi-supervised machine learning technique that automatically assigns a given document to a set of pre-defined categories based on its extracted features[8]. It provide comprehensive study on information retrieval accessibility, selection and management of large amounts of information on web that can be classify according to their category by applying supervised machine leaning algorithm namely NB, SVM, KNN and Decision Tree (DT). After different comparison the result shows each algorithm performance is depend on the characteristics of the datasets[9]. In[10] works on news text classification based on Latent Dirichlet Allocation (LDA) to reduce text dimension which is too high and get features by using topic model. Additionally, to solve multi-class of text problems Softmax regression algorithm uses a model's classifier. Proposed model achieved good classification results and effectively reduce the features dimension. In[11] due to large amount of unstructured data on internet meaningful information extraction is difficult to process by computers unless some effective and efficient techniques and algorithms are applied to reform data structure.

Proposed model is use for text mining (extracting meaningful information from text) in biomedical and health care domains with specific tasks and techniques including text pre-processing, classification and clustering. In[12] proposed a model for text classification based on Recurrent Neutral Network (RNN) as the acquisition function called Deep Active Learning (DAL). It uses internal state to process sequences of inputs due to this DAL no need preprocessing features extraction. Traditional Machine learning algorithm required less time to compute results after feature extraction in contrast DAL require much more time and need more labeled instances which gives high stable precision by only using 45% of the initial dataset. In[13] due to easily available of any type data Machine learning is capable to solve complex problems and enable automation in diverse domains. This study focus on the area of networking across different network technologies to address different problems for example traffic prediction, routing and classification, congestion control, resource and fault management, QoS and QoE management, and network security using Machine learning. Focus on Roman-Urdu data captured from different websites to analyze sentiments (positive/negative) comments/opinion from different people. Then compare SVM, LRSGD and NB supervised machine learning algorithm in which SVM produce 87.22% accuracy. Rashid, A. et al. Discuss the techniques of opinion mining which define as an intersection of computational linguistic and information retrieval which present in document[14]. Also cover long and short future area, challenges and gap in opinion mining discipline including the study on Supervised, Unsupervised machine learning as compared with case based reasoning techniques to perform computational treatment of sentiments. In[15] used movie review dataset by applying NB, Maximum Entropy and SVM learning techniques with different features i.e. POS, adjective, Unigram and Bigram to analyze document level sentiment. Proposed model gives best results which is 82.9% accuracy in case of SVM with Unigrams including three-fold cross validation method. In[16] proposed a model that uses 3-fold cross validation technique (involves partitioning of data into 3 subsets) on English language movie review for sentiment analysis. To train model it uses three major machine learning algorithms i.e. SVM, KNN and NB in which more than 80% accuracy achieved by NB and SVM than KNN on 800-1000 reviews.

As summarized in Gap Analysis Table 1 contains different researcher works that uses different supervised learning techniques which depend on training data to

predict class. However common challenges faced by these techniques that algorithm performance and accuracy depend on how data is mature and preprocess specially resource poorer languages i.e. (Roman-Urdu) in which proper linguistic and morphological structure is missing. Such limitations increase the probability of ML techniques to train and predict model and evaluate through precision and recall more accurately. As a remedy, some approaches of deep learning like Long Short Term Memory (LSTM)

Table 1. Summary of text classification techniques

Authors	Title	Dataset	Techniques	Limitation
In[1]	Sentiment classification of Roman-Urdu opinions using Naive Bayesian, Decision Tree and KNN	The blog contains public comments on "Effect of Facebook Usage" from (http://hamariweb.com/blogs/	Naive Bayesian outperformed Decision Tree and KNN	This work is only done on small dataset with two categories, 150 for each positive and negative.
In[13]	A compre-hensive survey on machine learning for networking: evolution, applications and research opportunities.	Waikato Internet Traffic Storage (WITS), UCI Knowledge Discovery Cyber-risk & Trust (IMPACT) Archive	Machine Learning for QoS and QoE management.	Hurdles in automatic network operation and management
In[4]	A survey on the state-of-the-art machine learning models in the context of NLP.	Wall Street Journal (WSJ) and Brown Corpus IRL Japanese Corpus of Holly Quran Etc.	HMM, CRF, maximum entropy (MaxEnt), SVM, Naïve Bays and deep learning.	In future the research community of NLP can increase the contribution of ML techniques for South East Asian languages too.
In[2]	Sentiment Analysis of Roman Urdu	Hamari web, youtube, drama on line, ytpak, facebook etc. 806 comments	Machine Learning Classifiers	-The longer sentences lose their structure during translation and some Urdu words cannot be converted by the tool to Roman script. - Small dataset with few categories
In[9]	Automatic text classification in information retrieval: A survey	Pima Indian Diabetes (768), Soybean (684), Vote (436), Blogger (100) total instances.	Naïve Bayes, SVM (SMO), KNN (IBK) and Decision Tree (J48).	Algorithm performance depends on the character-ristics of the datasets which affect the IR performance.
In[3]	Urdu language processing: a survey.	Becker–Riaz dataset The EMILLE dataset CLE dataset.	Part-of-speech (POS) · Named entity recognition (NER) · Sentence boundary detection (SBD)	Morphology is applied with different statistical techniques but not combine with any machine learning algorithm.
In[7]	Deep Learning-Based Sentiment Analysis for Roman Urdu Text.	Roman Urdu Dataset	Machine Learning Algorithm (NB, Random Forest, SVM and LSTM (Long Short Term Memory)	Work very efficient on sequential data models.

and Recurrent Neural Network (RNN) is use to precise results.

2. Proposed Methodology

In the methodology part, we start with the corpus collection which contains raw data that processes using preprocessing techniques for features selection to apply

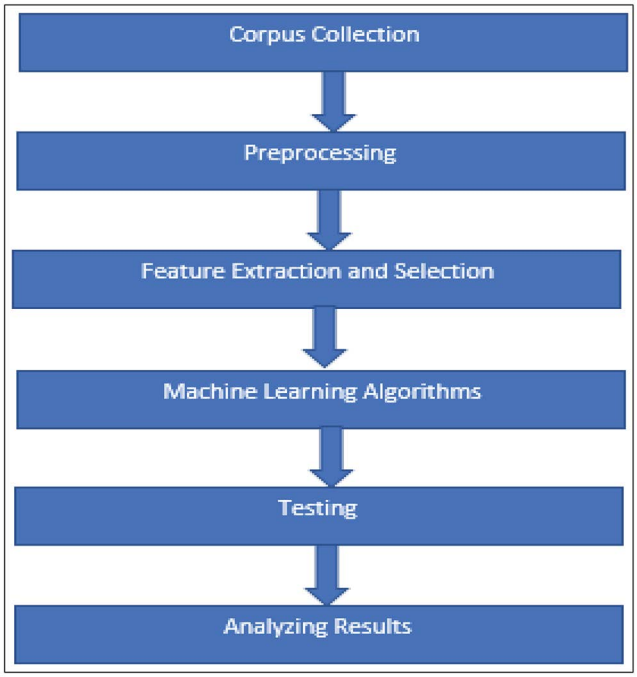

Figure 3. Summary of the process.

actual classification algorithms. The flow chart in Figure 3 summarizes the process which we followed for our technique.

3. Dataset Preparation

Supervised Machine Learning Algorithms need extensive amount of data to understand and predict. For this purpose, we collect data by using crawlers from different Roman-Urdu news agencies websites like, e.g., Jang News Roman-Urdu Page (https://jang.com.pk/roman). In total we collected 35000 sentences overall with so many other categories, but we selected top seven categories are as follows:

Accident, Sports, Weather, Arrest, Conference, Operation and Violence.

[0: 'Accident',
1: 'Operation',
2: 'Arrest',
3: 'Sports',
4: 'Violence',
5: 'Congres',
6: 'Weather']

Overall dataset is divided into two parts training and testing where training samples contain 80% and remaining 20% samples are testing.

3.1 Custom Stop-Words

Prepositions or those words which are not meaningful in nature is discarded in classification model for reducing processing in memory. Here, we create our custom define stop words list for Roman-Urdu which is useful to extract meaningful data for the classifiers few of them are mentioned below:

sw=["kia","ho","rahy","o'c","_","mai","gaya","ga","kis","mere","tum","nai"]

3.2 Feature Extraction and Selection

Feature selection is an important part of building machine learning models. We will be using the chi square test of independence to identify the important features.

'Accident':
. Most correlated unigrams:
. bus
hadsa
. Most correlated bigrams:
. keqareeb
traffichadsa
'Arrest':
. Most correlated unigrams:
giriftar
giraftar
. Most correlated bigrams:
giriftarpunjab
giraftarpunjab
'Congres':
. Most correlated unigrams:
sadarat
ijlas
. Most correlated bigrams:
ictislamabad
kaijlas
'Operation':
. Most correlated unigrams:

karwayi
operation
. Most correlated bigrams:
kikarwayi
search operation
'Sports':
. Most correlated unigrams:
cricket
pakistan
. Most correlated bigrams:
peshawarzalmi
world cup
'Violence':
Most correlated unigrams:
qatal
firing
. Most correlated bigrams:
ki firing
par firing
'Weather':
. Most correlated unigrams:
garmi
barish
. Most correlated bigrams:
punjabke
barishpunjab

3.3 Term Frequency- Inverse Document Frequency

Specifically, for each term in our dataset, we will calculate a measure called Term Frequency, Inverse Document Frequency, abbreviated to tf-idf.

4. Results and Discussion

Figure 4 shows experiment results of various Machine Learning Algorithms such as K-Neighbors, Linear-SVC, Logistic Regression, Multinomial NB, Nearest Centroid, Passive Aggressive Classifier, Perceptron, Random Forest Classifier, Ridge Classifier, and SGD Classifier. The result shows accuracies from 82.20% to 93.06% except Random Forest model which gives only 34.08% as shown in Table 1. We observed SGD classifier gives better results 93.50% as compare to top 5 algorithms with different variations in all categories according to their applied techniques. Here SDG classifier has significant impact than others which give above 90% of measured matrix than other classifiers which identify few categories above the mark of 90%. According to given matrix it's not necessary that all model performed well, we explain few model matrices in which results can be easily understand. Our model is dividing dataset into training and testing which leads to analyses main sources of misclassification on the test set. Major source to identify error is confusion matrix based on predicted and actual labels discrepancies. Figure 5 Linear-SVC Confusion Matrix shows a correct prediction on diagonal side where correct label of accident category is 486, operation 544, arrest 399, sports 659, violence 1057, congress 413 and weather 297. However, Figure 6

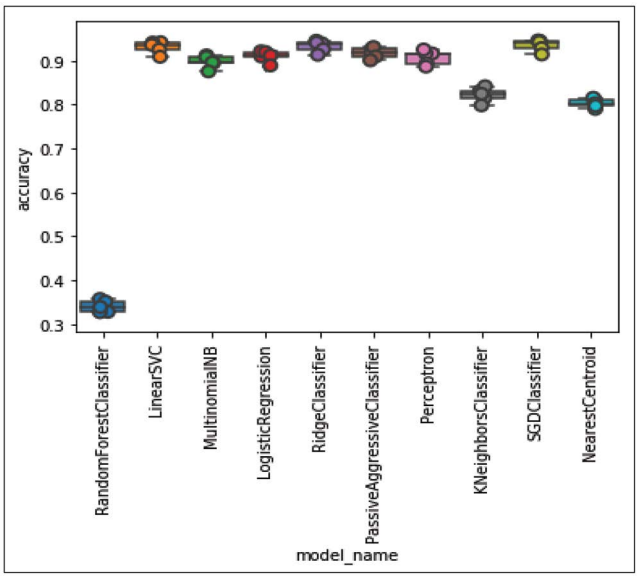

Figure 4. Algorithm accuracy.

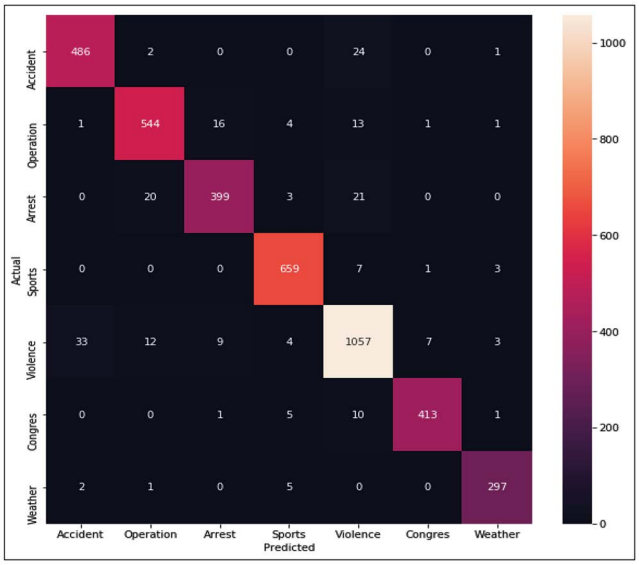

Figure 5. Linear-SVC Confusion Matrix.

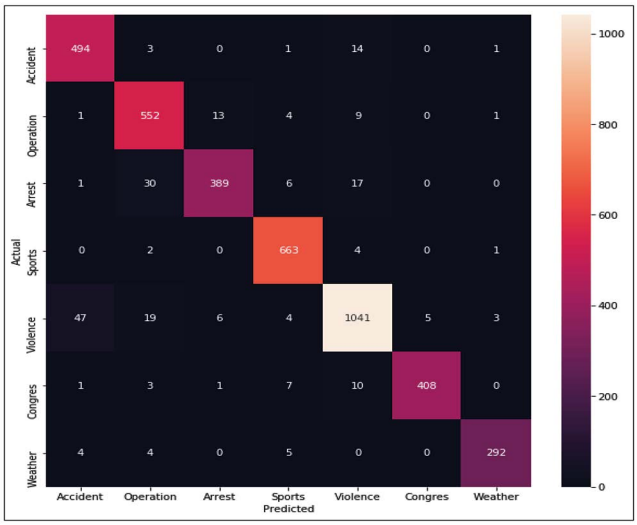

Figure 6. GD Classifier Confusion Matrix.

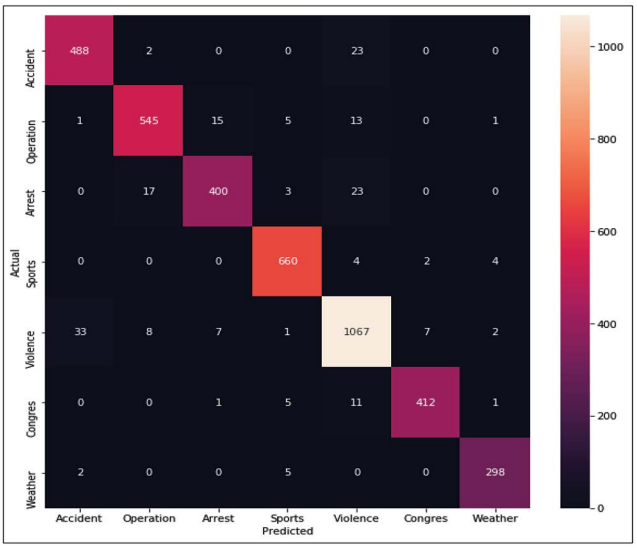

Figure 7. Ridge Classifier Confusion Matrix.

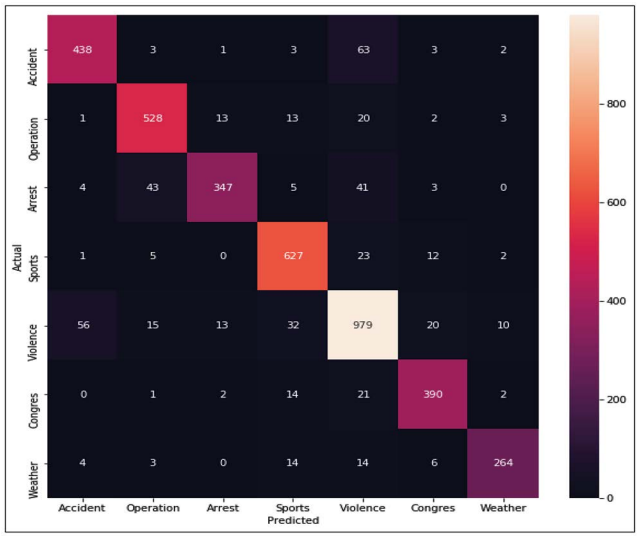

Figure 8. Random Forest Confusion Matrix.

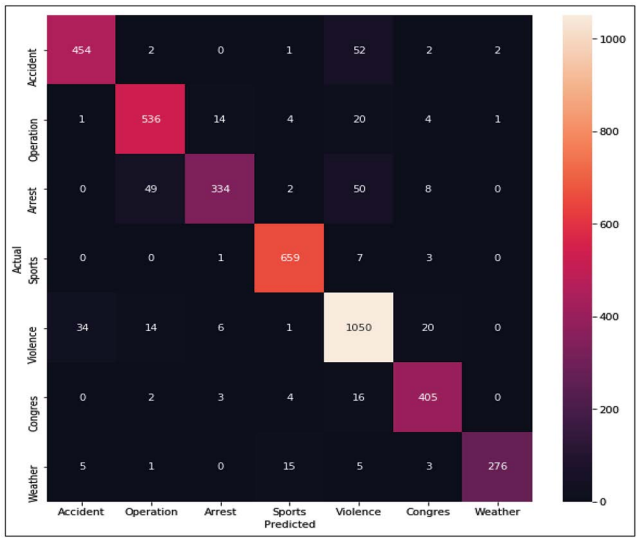

Figure 9. Multinomial NB Confusion Matrix.

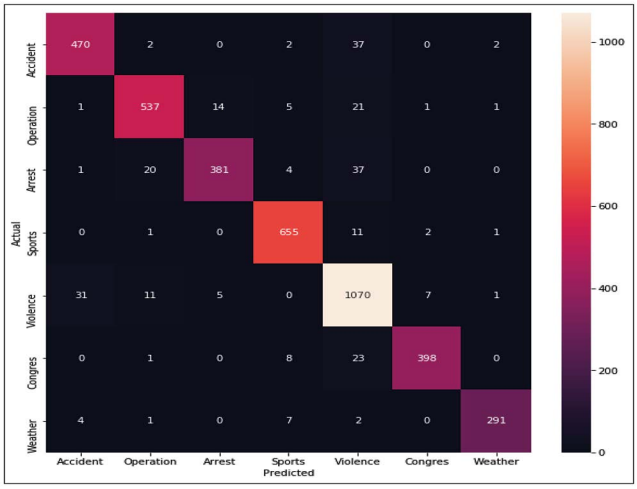

Figure 10. Logistic regression confusion Matrix.

demonstrates the correct prediction of seven categories where accident category is 494, operation 552, arrest 389, sports 663, violence 1041, congress 408 and weather 292. Further Figure 7 describe actual and predicted matrix in which accident, operation, arrest, sports, violence, congress and weather (488, 545, 400, 660, 1067,412 and 298). Similarly, Figure 8-10 Confusion Matrix shows true labels diagonal above mentioned category.

After training our model we test few data that give excellent results.

"DarjaHararat 44 Degree Tak Janay KaImkaan"
- Predicted as: 'Weather'

"Aj England and Pakistan ki teams maibohatachamuqablahuha"
- Predicted as: 'Sports'

"Karachi maizordardhamaka"
- Predicted as: 'Violence'

"Karachi maiajTaizHawaonchalin"
- Predicted as: 'Weather'

"Karachi mai traffic hadsa 3 log mar gaye"
- Predicted as: 'Accident'

"AjjWazir-e-AzamIjlazkarain gain"
- Predicted as: 'Congres'

"Na maloomafradkitarafsai firing"
- Predicted as: 'Violence'

5. Conclusion

This study introduced model for news headline text classification on Roman-Urdu Language by taking data from different websites using web scrapper tool. In this work, the dataset is divided into two parts in which 80% is dedicated for training and remaining is used for testing. Our system implements Unigram and Bigram language models for identify long distance dependencies between sentences. Moreover, to analysis the feature from sentences TF-IDF and Count Vector information retrieval models have been used. We conducted comprehensive experiments on roman news classification where various machine learning techniques are used to train system. To analyze the results different evaluation matrices are used like Precision, Recall, F1-score, Confusion Matrix and Accuracy. It found that SGD classifier extract more features comparatively others. Accuracy of SGD classifiers 94% and others classifiers is less 93%. In future, we will extend more classes and use other dataset in Urdu and English and will be compared with all.

6. Future Work

In future, this work shall be extended in order to implement deep learning methods.

7. References

1. Bilal M. Sentiment classification of Roman-Urdu opinions using Naive Bayesian, Decision Tree and KNN classification techniques, Journal of King Saud University - Computer and Information Sciences. 2016; 28(3):330–44. https://doi.org/10.1016/j.jksuci.2015.11.003.
2. Ayesha R. Sentiment analysis for Roman Urdu, Mehran University Research Journal of Engineering and Technology. 2019; 38(2):463–70.
3. Daud A, Khan W, Che D. Urdu language processing: A survey, Artificial Intelligence Review. 2017; 47(3):279–311. https://doi.org/10.1007/s10462-016-9482-x.
4. A survey on the state-of-the-art machine learning models in the context of NLP. Date accessed: 10/2016. https://www.researchgate.net/publication/311436163_A_survey_on_the_state-of-the-art_machine_learning_models_in_the_context_of_NLP.
5. Sequence to Sequence Networks for Roman-Urdu to Urdu Transliteration. Date accessed: 08/12/2017. https://arxiv.org/abs/1712.02959.
6. Urdu Text Classification using Majority Voting. Date accessed: 2016. https://thesai.org/Publications/ViewPaper?Volume=7&Issue=8&Code=IJACSA&SerialNo=36, https://doi.org/10.14569/IJACSA.2016.070836.
7. Ghulam H. Deep learning-based sentiment analysis for Roman Urdu text, Procedia Computer Science. 2019; 147:131–35. https://doi.org/10.1016/j.procs.2019.01.202.
8. Dalal MK. Automatic text classification: A technical review, International Journal of Computer Applications. 2011; 28(2):37–40. https://doi.org/10.5120/3358-4633.
9. Automatic text classification in information retrieval: A survey. Date accessed: 03/2016. https://www.researchgate.net/publication/307436499_Automatic_Text_Classification_in_Information_retrieval_A_Survey.
10. News text classification model based on topic model. Date accessed: 26-06/2016. https://ieeexplore.ieee.org/document/7550929.
11. A brief survey of text mining: Classification, clustering and extraction techniques. Date accessed: 2016. https://arxiv.org/pdf/1707.02919.pdf.
12. Deep Active Learning for Text Classification. Date accessed: 2018. https://dl.acm.org/citation.cfm?id=3271578.
13. Boutaba R. A comprehensive survey on machine learning for networking: Evolution, applications and research oppor-

tunities, Journal of Internet Services and Applications. 2018; 9(16):1–16. https://doi.org/10.1186/s13174-018-0087-2.

14. Rashid A. A survey paper: Areas, techniques and challenges of opinion mining, International Journal Computer Science. 2013; 10(6):18–31.
15. Pang B, Lee L, Vaithyanathan S. Thumbs Up? Sentiment Classification using Machine Learning Techniques. Proceedings of ACM-ACL Conference on Empirical Methods in Natural Language Processing; 2002. p. 79–86. https://doi.org/10.3115/1118693.1118704.
16. Kalaivani P, Shunmuganathan D. Sentiment classification of movie reviews by supervised machine learning approaches, Indian Journal of Computer Science and Engineering. 2013; 4(4):285–92.