TUNING LANGUAGE PROCESSING APPROACHES FOR PASHTO TEXTS CLASSIFICATION

Jawid Ahmad Baktash, Mursal Dawodi

LIA, Avignon, University Avignon, France

ABSTRACT

Nowadays, text classification for different purposes becomes a basic task for concerned people. Hence, much research has been done to develop automatic text classification for the majority of national and international languages. However, the need for an automated text classification system for local languages is felt. The main purpose of this study is to establish a novel automatic classification system of Pashto text. In order to follow this up, we established a collection of Pashto documents and constructed the dataset. In addition, this study includes several models containing statistical techniques and neural network neural machine learning including DistilBERT-base-multilingual-cased, Multilayer Perceptron, Support Vector Machine, K Nearest Neighbor, decision tree, Gaussian naïve Bayes, multinomial naïve Bayes, random forest, and logistic regression to discover the most effective approach. Moreover, this investigation evaluates two different feature extraction methods including bag of words, and Term Frequency Inverse Document Frequency. Subsequently, this research obtained an average testing accuracy rate of 94% using the MLP classification algorithm and TFIDF feature extraction method in single label multi-class classification. Similarly, MLP+TFIDF with F1-measure of 0.81 showed the best result. Experiments on the use of pre-trained language representation models (such as DistilBERT) for classifying Pashto texts show that we need a specific tokenizer for a particular language to obtain reasonable results.

KEYWORDS

Pashto, DistilBERT, BERT, Multi-lingual BERT, Multi-layer Perceptron, Support Vector Machine, K Nearest Neighbor, Decision Tree, Random Forest, Logistic Regression, Gaussian Naïve Bayes, Multinomial Naïve Bayes, TFIDF, Unigram, Deep Neural Network, Classification

1. INTRODUCTION

The evolution of technology instigated the existence of an overwhelming number of electronic documents therefore text mining becomes a crucial task. Many businesses and individuals use machine learning techniques to classify documents accurately and quickly. On the other hand, more than 80% of organization information is in electronic format including news, email, data about users, reports, etc. (Raghavan, 2004). Text mining attracted the attention of researchers to automatically figure out the patterns of millions of electronic texts. Among other opportunities, this provides a facility for users to discover the most desirable text/document.

Pashto is a resource poor language and the unavailability of standard, public, free of cost datasets of text documents is a major obstacle for Pashto's document classification. Automatic text document classification and comparatively analyze the performance of different models are the main gaps in Pashto text mining. This research is the first attempt to classify Pashto documents into eight classes including Sport, History, Health, Scientific, Cultural, Economic, Political, and Technology. Besides, this is the initial and novel work on Pashto text multi-label classification.

David C. Wyld et al. (Eds): DBML, CITE, IOTBC, EDUPAN, NLPAI -2023
pp.297-311, 2023. IJCI – 2023 DOI:10.5121/ijci.2023.120222

The main contributions of this research are to:

- Designing two Pashto document dataset and make them publicly and free of cost available in the future.
- Compare the performance of 32 distinct models on Pashto text single label and multilabel classification.
- Evaluate the performance of standard pre-trained language representation model (DistilBERT) on Pashto language processing

2. PASHTO LANGUAGE

Pashto is an Iranian language, a branch of the Indo-European language family, spoken natively by a majority of Afghans, more than seven million Pakistani, and 5000 Iranian (Tegey and Robson 1996). Pashtuns, people whose mother tongue is Pashto, usually live in the south of Afghanistan and north of Pakistan. This language has three main distinct dialects based on the geographic location of native Pashtun residents. The diversity of dialects even effects on spelling of Pashto text since some speakers pronounce the “sh” like “x” in Greek or “ch” in Germany rather than “sh” in English (Tegey and Robson 1996). Besides, no transliteration standard exists for rendering the Pashto text to the roman alphabet and that is why both Pashto and Pashtu are the correct spelling form (Tegey and Robson 1996). However, one can find some official recommendations relevant to Pashto writing and speaking. Moreover, it does not have any standard rules for writing and pronunciation therefore the authors often write one word in several ways and the speakers pronounce them in various ways (Tegey and Robson 1996). The representation of letters in this language is similar to Arabic and Persian with some extra characters. Fig 1 demonstrates the alphabet representation in the Pashto language.

Pashto differentiate nouns based on genders and distinguishes the form of verbs and pronouns for masculine and feminine nouns, as an example, دا د هغې مور ده (daa de haghe mor da) means she is her mother and دا د هغهپلار دی (daa de hagha pelar de) indicates he is his father. Morphemes like plural morphemes in Pashto added another challenge to this language (Kamal et al., 2016), e.g. the plural form of زو ی (son) is زامن (zaamen, sons) while كتابونه (ketaboona, books) is the plural form of كتاب (ketab, book) and the plural form of انجلی (enjeley, girl) is انجونې (anjoone, girls). Besides, news, articles, and other online and offline scripts are not written/typed by Pashto native speakers hence the probability of grammar and spelling error is high (Tegey and Robson 1996). Additionally, grammar in Pashto is not as traditional as other Indo-European languages. Although nowadays several Pashto grammar books are published. Still, they have contradicted each other in some parts (Tegey and Robson 1996). Furthermore, other languages spoken in the vicinity of Pashtun areas have major influences on this language that caused arriving of foreign words in Pashto for instance. some Pashtuns combine Urdu or Dari words with Pashto while speaking or in their written text.

3. RELATED WORKS

Many studies on document classification have already been conducted in international and western languages. As a recent work in text document classification, Gutiérrez et al. (2020) developed a COVID 19 document classification system. They compared several algorithms including SVM, LSTM, LSTMreg, Logistic Regression, XML-CNN, KimCNN, BERTbase, BERTlarg, Longformer, and BioBERT. The best performance was achieved by BioBERT with an accuracy of 75.2% and a micro-F1-measure of 0.862 on the test set. In recent years some researchers started to work on document classification in Asian and local languages. Ghasemi S. and Jadidinejad A.H. (Ghasemi & Jadidinejad, 2018) used character level convolutional neural network to classify Persian documents. They obtained 49% accuracy which was much higher compared to the results of Naïve Bayes and SVM. Similarly, Baygin M. (Baygin, 2018) used the Naïve Bayes method and ngram features to classify documents in Turkey into economic, health, sports, political, and magazine newsgroups. They performed their proposed model on 1150 documents written in Turkey. The best performance was achieved by the 3-gram technique with 97% accuracy on sport, politics, and health documents, 98% on a magazine, and 94% on economic documents.

Similarly, Pervez et al. (2020) obtained impressive results using a single layer convolutional neural network with different kernel sizes to classify Urdu documents. They evaluated the model on three different Urdu datasets including NPUU, naïve, and COUNTER. NPUU corpus consists of sport, economics, environment, business, crime, politics, and science and technology Urdu documents. Likewise, naïve contains Urdu documents related to sports, politics, entertainment, and economic. Finally, the main document classes in COUNTER dataset are business, showbiz, sports, foreign, and national. Consequently, they obtained 95.1%, 91.4%, and 90.1% accuracy on naïve, the COUNTER, and NPUU datasets, respectively. Pal, K., & Patel, Biraj. V. (2020). Automatic Multiclass Document Classification of Hindi Poems using Machine Learning Techniques. 2020 International Conference for Emerging Technology (INCET), 1–5. https://doi.org/10.1109/INCET49848.2020.9154001 categorized Indi poem documents into three classes romance, heroic, and pity according to the purpose of the poem. They evaluated several machine learning techniques. The maximum accuracy 56%, 54%, 44%, 64%, and 52% using Random Forest, KNN, Decision Tree, Multinomial Naïve Bayes, SVM, and Gausian Naïve Bayes.

Some researches were conducted in the context of multi-label classification of articles, recently. As a similar work (Elnagar, 2020), constructed two separate large corpora for single label and multi-label Arabic news categorization. They evaluated the performance of several deep learning algorithms on classifying Arabic articles. Finally, the best performance of 96.94% accuracy and 88.68% overall accuracy using attention-GRU in the context of single and multi-label classification process, respectively. Similarly, Al-Salemi et al. (2018) introduced a new Arabic multi-label benchmark dataset named “RTANews”. Next, they examine the performance of four problem transformation-based approaches, including Binary Relevance, Classifier Chains, Calibrated Ranking by Pairwise Comparison, and Label Powerset, and five algorithm adaptation-based techniques, including Multi-label K-Nearest Neighbors, Instance-Based Learning by Logistic Regression Multilabel, Binary Relevance KNN, and RFBoost. As a result, the transformation approaches were performed better with SVM as a base classifier, and the best performance was achieved with RFBoost method. Likewise, Qadi et al. (2019) established an Arabic multi-label dataset with four main categories: Business, Sports, Technology, and the Middle East. They utilized Logistic Regression, Nearest Centroid, DT, SVM, KNN, XGBoost, Random Forest Classifier, Multinomial, Ada-Boost, and MLP to determine relevant labels for Arabic news. They claimed that SVM with 97.6% accuracy outperformed other methods.

Recent studies utilized transformer based pre-trained models. Tokgoz et al. (2021) used BERT and DistilBERT with different tokenizers including Turkish tokenizer to classify news in Turkish language. DistilBERT with Turkish tokenizer obtained the best performance with 97.4% accuracy. Similarly, the study by Gutiérrez et al. (2020) compared the performance of traditional machine learning methods, convolutional neural models, and pretrained language models on COVID 19 document classification. As a result, the reasonable accuracy of 75.2% and 74.4% achieved with BioBERT and BERT large respectively. Correspondingly, Farahani et al. (2021) developed a monolingual transformer-based model for Persian language and evaluated the proposed model in several datasets. Finally, the investigation declares that the ParsBBERT outperforming both multilingual BERT and other prior works in Persian down-stream NLP tasks such as Sentiment Analysis, Text Classification and Named Entity Recognition. Dai and Liu (2020) used BERT model for multilabel classification of Chinese Judicial documents.

As of history, there is no documented classification for the Pashto language. The only work to classify the Pashto text, which in some respects relates to our work, was done by S. Zahoor et al. (2020). They have developed a character recognition system that captures images of Pashto letters and automatically classifies them by predicting a single character.

4. CORPORA

In this study we constructed two datasets corresponding to multilabel and single label multiclass document classification.

4.1. Single Label Dataset

This research gathered 800 manuscripts from several online books, articles, and web pages to make a corpus for text document classification analysis. Subsequently, we manually assigned label/s by setting a number to every single document in relevance to the category it belongs. We collected 100 Pashto documents for each class including history, technology, sport, cultural, economic, health, political, and scientific. Besides, we increased the dataset with 475 news-related documents.

4.2. Multi Label Dataset

The structure of the corpus is altered compared to single-label document corpus. Each document was assigned to multiple related classes. Here, we considered the news category along with the prior labels. The average length of the documents is 3119 words where the shortest document has 232 words and the longest one includes 31740 words. Similarly, the total number of words is 3289214 in the single label and 3976976 words in multilabel datasets. The average number of labels per document is 2.5.

In the next step, the authors preprocessed the dataset by applying some spelling and grammar modification, removing any noisy and senseless symbols including non-language characters, special symbols, numeric values, and URLs. As a result, we standardized and normalized the texts within the documents.

5. MODELS

This section details the methods followed to accomplish the study for categorizing Pashto sentences and documents. We used three different types of classifiers with different feature extraction and tokenization methods.

5.1. Traditional Models

As mentioned in prior sections, this work observed different classifiers methods including Naïve Bayes, Multinomial Naïve Bayes, Nearest Neighbor, Random Forest, Decision Tree, SVM, Logistic Regression. This project fine-tuned the value of K finally the result shows that the optimum value for k occurs at k=5.

5.2. Neural Network Models

We used Multilayer Perceptron Network (MLP) as the neural network model. MLP is a neural network classifier which is a subset of machine learning consists of neurons and layers. MLP consists of several layers including one input, one output, and hidden layers. The outputs of one perceptron are fed as input to subsequent perceptron. This experiment used a single hidden layer with 20 neurons. We used backpropagation and gradient descent to provide the ability to propagate errors back to earlier layers. Moreover, we shuffled the samples to reduce noise by feeding different inputs to neurons in each iteration and as a result, make good generalizations. The activation function used in this model is Rectified Linear Unit (ReLU) function, which is a non-linear activation function, to decrease the chance of vanishing gradient. ReLU is defined in equation 1 where f(a) is considered to be zero for all negative numbers of a. Finally, we used Adam as an optimization algorithm.

$$f(a) = \max(0, a) where\ a = Wx + b \quad (1)$$

5.3. Contextual Word Representation Models

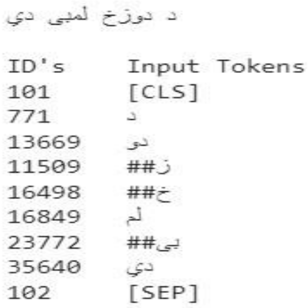
Nowadays, contextual word representation models such as BERT and DistilBERT are used in varied tasks of NLP including text classification. BERT was initially developed by Google AI as base and large pre-trained models. The difference between two models is on the number of transformer blocks. Base model uses 12 layers encoders where the large BERT utilizes 24 layers of encoder on top of each other. Thus, the base model contains 12 self-attention heads with hidden size of 786. The maximum number of tokens handled by base model is 512. It contains some default tokens as [CLS] that marks the starting of the segment and [SEP] that differentiates the segments. DistilBERT is faster and smaller distilled version of BERT which is more suitable in NLP tasks. The architecture of DistilBERT contains 6 layers, 12 heads, and 786 dimensions.

We used BERT-base-multilingual-cased and DistilBERT-base-multilingual-cased models with the tokenizer of BERT-base-multilingual-cased and DistilBERT-base-multilingual-cased, respectively. In our model the final hidden state of the first token [CLS] demonstrates the entire sequence. An activation function classifier on the of the model is used to predict the related class of a text document.

5.4. Tokenization

Tokenization shrinks the sentences into lexicons/tokens (e.g. ['واه', 'ډیر', 'ښکلی']) using available token list. The special tokens specify the start and end of the sequence. We specified the maximum length of tokens. Hence, the extra tokens are discarded if the sequence is longer than the maximum size while extra empty tokens are added to shorter sequences.

In this study we tokenized the text by using NLTK work tokenizer in the traditional and MLP models. On the other hand, the standard tokenization process was used in BERT and BistilBERT models. The procurement root of separate tokens in Pashto is a more challenging task due to morphemes and other issues in Pashto literature. Thus, this work used lexicons in their default forms. Figure 1 represents how a sentence in Pashto is tokenized into tokens each containing an ID.



```
د دوزخ لمبی دي

ID's     Input Tokens
101      [CLS]
771      د
13669    دو
11509    ##ز
16498    ##خ
16849    لم
23772    ##بی
35640    دي
102      [SEP]
```

Figure 1 Pashto Sentence Tokenization using standard DistilBERT Tokenizer

6. Evaluation Matrices

This study evaluates the performance of different classifiers using separate feature extraction methods. We considered four metrics Precision (equation 2), Recall (equation 3), F1-measure (equation 4), and accuracy (equation 5) to analyze the outcome of different models of the first and second group used in this experiment. Precision, which is called positive predictive values, is the percentage of examples that the classifier predicts accurately from the total samples predicted for a given tag. On the other hand, Recall which is also referred to as sensitivity determines the percentage of samples that the classifier predicts for a given label from the total number of samples that should be predicted for that label. Accuracy represents the performance of the model while is referred to the percentage of texts that are predicted with the correct label. We used F1-measure to measure the average between Precision and Recall values. There are mainly four actual classes true real positive (TP), false real positive (FP), true real negative (TN), and false real negative (FN). TP and TN are the accurate predictions while FP and FN are related to imprecise estimations:

$$\text{True class} = \{TP_1, TP_2, \ldots, TP_n, \} \cup \{TP_1, TP_2, \ldots, TP_n, \}$$

$$\text{False class} = \{FP\_1, FP\_2, \ldots, FP\_n, \} \cup \{FP\_1, FP\_2, \ldots, FP\_n, \}$$

$$Precision = \llbracket TP \rrbracket \ \_i/(\llbracket TP \rrbracket \ \_i + \llbracket FP \rrbracket \ \_i) \quad (2)$$

$$Recall = \llbracket TP \rrbracket \ \_i/(\llbracket TP \rrbracket \ \_i + \llbracket FN \rrbracket \ \_i) \quad (3)$$

$$F1 - measure = 2PR/(P + R) \quad (4)$$

$$Accuracy = (\llbracket TP \rrbracket \ \_i + \llbracket TN \rrbracket \ \_i)/(\llbracket TP \rrbracket \ \_i + \llbracket FP \rrbracket \ \_i + \llbracket TN \rrbracket \ \_i + \llbracket FN \rrbracket \ \_i) \quad (5)$$

Consequently, this study computes weighted average values for Precision, Recall, and F1-measure of all classes to compare the efficiency of each technique. In multi-label classification, the weighted average evaluates metrics for all labels and calculates their averages weighted by the total number of true instances per each label. Besides, to evaluate the performance of individual algorithms in classifying documents into multiple tags, we used AUC (area under the curve) along with other criteria. AUC is a classification threshold invariant which, means that it calculates the performance of the models concerning all possible score thresholds, regardless of the importance of each threshold. It corresponds to the array of samples and classes. The probability estimates are related to the probability of the class that has a larger label per output of the classifier. To compute AUC-ROC we used the ROC_AUC_score method of Python scikit learn metrics library. Additionally, we evaluated the sample average scores for precision, recall and f1-measure. The sample average estimates metrics per instance then averages the results.

For analysis of the third model, we used accuracy and loss metrics in single label classification task. On the other hand, the authors experiment the performance of third group of models on Pashto text classification by using hamming score and hamming loss. Hamming score corresponds to the portion of accurate predictions associated to the overall labels while the fraction of wrong labels to the entire number of labels indicates the hamming loss.

7. RESULTS AND DISCUSSION

This work was implemented in python 3.6.9 using a computer in windows 10 environment with Intel core i7(TM) 2.80 GHz 2.90 GHz processor. TensorFlow version 2.3.1 and Keras version 2.4.3 were used for implementing the diverse classification models. Although, this study is not able to outperform recent related research such as (Elnagar, 2020) undoubtedly it is a valuable achievement in the field of classification of Pashto texts as no research has been done on this subject so far.

This study has some limitations due to the immature context of the Pashto language. There is not any special toolkit for processing Pashto language like Hazm for Persian language and NLTK for English language. The dataset used in this study is very short with only 800 records. Additionally, this experiment only considered 8 separate classes for Pashto documents. However, our future goal is to expand the corpus and use more hybrid algorithms to achieve better performance.

7.1. Multiclass Single Label Classification Using Model I and Model II

MLP with unigram feature extraction technique illustrated the best performances among others with the gained average accuracy of 94%. Besides, it obtained 0.94 as weighted average Precision, Recall, and F1-measure scores. As one can see in Figure 8, the maximum weighted average Precision using MLP and unigram is 0.91. The obtained results are presented in Table 7.1 and Figures 4 to 6. Table 7.1 demonstrates obtained accuracy using different techniques. Similarly, figure 2 denotes the F-measure weighted average values obtained when testing distinct techniques.

Multinomial Naïve Bayes with Unigram achieved 88% accuracy while it decreased by 7% replacing Unigram with TFIDF which indicates that it performed better with Unigram text embedding technique. However, Gaussian Naïve Bayes obtained 87% accuracy using TFIDF vector representations which are 11% higher compared to Gaussian Naïve Bayes +Unigram. Even though, Gaussian Naïve Bayes has an impressive result of 0.85 as weighted Precision result, but it obtained F1-measure of only 0.77 due to its low Recall score of 0.76. In contrast to Gaussian Naïve Bayes, Decision Tree obtains 5% more accuracy using Unigram rather than TFIDF. Performance of Logistic Regression, SVM, Random Forest, and KNN with both TFIDF and Unigram are comparable with only 1% change in accuracy and 0-0.2 variation in F1-measures.

The combination of SVM and unigram represented 84% average accuracy while this value is reduced by 1% using TFIDF. Therefore, similar to several classification studies SVM performed good in Pashto text document classification. In contrast to the work by Mohtashami and Bazrafkan (2014), KNN attained only 71% as average accuracy using TFIDF method that is decreased to 70% after altering the feature extraction method from TFIDF to Unigram. The least performance belongs to Decision Tree method with TFIDF technique in this comparison experiment which is only 64% accuracy. This method also has low performance (with F1-

measure of 0.69) using unigram extraction method. On the other hand, the entire methods performed their worst with bigram technique as illustrated in Table 1.

Table 1 Average accuracy using different classification and feature extraction techniques

Technique	Feature Extraction Method	Accuracy
Gaussian Naïve Bayes	Unigram	0.76
	TFIDF	0.87
	Bigram	0.72
Multinomial Naïve Bayes	Unigram	0.88
	TFIDF	0.81
	Bigram	0.78
Decision Tree	Unigram	0.69
	TFIDF	0.64
	Bigram	0.44
Random Forest	Unigram	0.82
	TFIDF	0.81
	Bigram	0.67
Logistic Regression	Unigram	0.85
	TFIDF	0.84
	Bigram	0.36
SVM	Unigram	0.83
	TFIDF	0.84
	Bigram	0.65
K Nearest Neighbor	Unigram	0.7
	TFIDF	0.71
	Bigram	0.31
Multilayer Perceptron	Unigram	0.91
	TFIDF	0.94
	Bigram	0.88

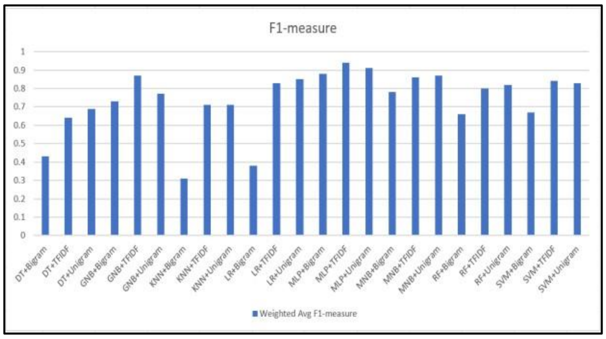


Figure 2 Weighted average F1-measure For Single Label Classification

The outcome of each model is different according to the separate class label. As an example, KNN employed unigram has 0.98 F1-measure related to History tag. However, it obtained only 0.37 for scientific documents.

Similarly, all models illustrated good F1-measure for documents relevant to History except Gaussian Naïve Bayes. DT achieved high F1-measure only by predicting documents related to History. Experiments show that MLP models and the combined model of Random Forest with Unigram more accurately predicts cultural documents compared to other models. MLP with TFIDF and Gaussian Naïve Bayes with Unigram with 0.95 and 0.93 F1-measure have the most accurate Economic class predictions in this experiment. On the other hand, the implementation of Gaussian Naïve Bayes and SVM with

Unigram represents the most precise results in the context of health documents. Similarly, MLP with TFIDF obtains the highest F1-measure of 1 on predicting Politic documents. All models failed to predict scientific documents precisely except MLP + TFIDF model with F1-measure of 0.89. MLP with Unigram with F1measure 0.98 best performed in discovering texts related to Sport class. Similarly, Random Forest and Gaussian Naïve Bayes with TFIDF came in second with F1-measure 0.95 in this era. Likewise, MLP+TFIDF and Gaussian Naïve Bayes + TFIDF models best predict texts belonging to Technology class with F1-measure 1 and 0.97 respectively.

7.2. Multiclass Multi Label Classification Using Model I and Model II

In multilabel classification, MLP technique illustrated the best performance similar to the single label classification with sample average F1-measure of 0.81 using TFIDF technique. Despite, the MNB+Bigram technique has the highest AUC score of 85.7% but its F1-measure is 3% lower than MLP+TFIDF. Afterward, SVM+TFIDF and SVM+ Unigram obtained sample average F1-measure of 0.74. On the other hand, according to the Precision metric the highest value of 0.86 achieved using SVM+TFIDF technique. Using any of the MNB+Bigram, MLP+TFIDF. However, Technology related documents were better detected using MLP+Bigram technique. With SVM+Unigram, MLP+Unigram algorithms the AUC scores (figure 4) are higher than 80%. The least performance obtained using LR+Bigram and LR+Trigram models.

Table 2 depicts the average accuracy obtained using different algorithms according to separate labels. Regarding separate label the highest prediction accuracy related to History, Culture, and Economics achieved by MLP+TFIDF. However, the SVM+TFIDF algorithm presented the best performance based on News, Health, and Politic label groups. On the other hand, the GNB+TFIDF predicted Scientific related documents more accurately.

Table 2 Obtained accuracy related to separate labels in multi-label classification

Model	H	C	E	H	P	Sc	Sp	T	N
DT+B	90	65	76	82	62	81	96	91	74
DT+T	88	72	81	81	77	83	94	92	81
DT+U	91	71	78	84	72	79	92	92	86
GNB+	87	72	83	89	82	86	98	95	84
GNB+T	94	78	89	92	84	94	94	93	84
GNB+Tri	89	73	82	87	80	83	96	92	90
GNB+U	93	65	86	88	80	86	96	88	80
KNN+B	91	63	79	85	65	82	94	92	76
KNN+T	92	69	84	90	78	83	95	89	81
KNN+U	93	72	83	88	80	88	96	90	85
LR+B	89	69	84	80	58	81	94	92	64
LR+T	91	70	84	80	80	86	95	90	90
LR+U	92	75	82	82	84	82	92	90	90
MLP+B	95	78	89	89	79	90	96	96	88
MLP+T	96	86	92	93	89	92	96	95	94
MLP+U	90	83	87	94	86	91	98	95	91
MNB+B	92	78	8	87	86	89	94	91	92
MNB+T	90	70	84	83	86	87	93	90	90
MNB+Tri	89	73	8	83	76	87	92	88	90
MNB+U	93	65	86	88	80	86	96	88	80
RF+B	921	733	835	898	733	858	976	99	851
RF+T	93	77	87	90	82	85	97	94	90
RF+U	93	78	86	93	85	92	99	92	89
SVM+B	91	69	84	86	74	83	94	94	85
SVM+T	91	80	88	94	91	88	97	93	95
SVM+Tri	96	70	83	85	61	80	95	92	75
SVM+Tri	96	70	83	85	61	80	95	92	75
SVM+U	92	76	87	92	83	89	98	94	86
DT+Tri	93	61	81	85	60	81	97	92	68
LR+Tri	95	68	82	81	58	77	90	90	67
KNN+Tri	91	67	87	26	58	81	91	93	68
MLP+Tri	98	68	82	83	66	80	95	92	76
RF+Tri	91	63	81	85	58	87	94	93	80

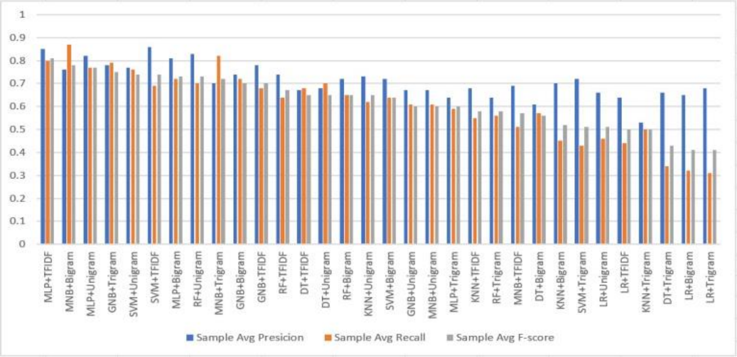


Figure 3 Sample average Precision, Recall, and F1measure in multi-label classifier

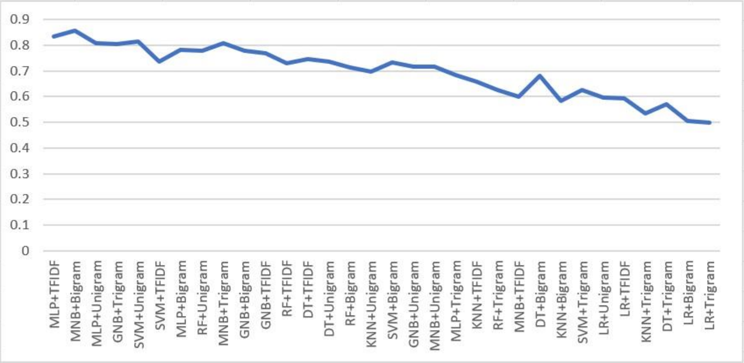


Figure 4 Compression of AUC result based on various algorithms

Similarly, RF+Unigram more precisely determined sport documents. Table 3 represents the weighted average Precision, Recall, F1-measure, and Support metrics corresponding to multi-label Pashto article classification. As one can see the MLP+TFIDF technique achieves the highest weighted average Precision.

As represented in tables 13 and 14, some diverse methods have different performances with variant feature extraction techniques. For example, the GNB works the best using trigram method, however, SVM outperformed with unigram. Similarly, the combination of MLP with TFIDF and MNB along with bigram performance more precisely based on this study.

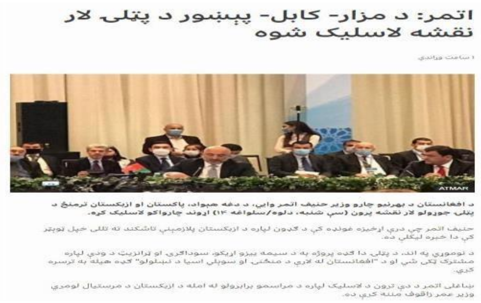
MLP+TFIDF technique achieves the highest weighted average Precision. Some diverse methods have different performances with variant feature extraction techniques. For example, the GNB works the best using trigram method, however, SVM outperformed with unigram. Similarly, the combination of MLP with TFIDF and MNB along with bigram performance more precisely based on this study.

The multi-label classification models predict the exact tags for an article. Figures 5 and 6 demonstrate three news articles from the BBC Pashto News website and the predicted labels. The true labels for figure 5 are political, economic, and news. Similarly, the news article represented in figure 7.6 is related to sport and the last figure is the news about COVID-19 and flights between Saudi Arabia and some other countries. Fortunately, our model predicts all tags accurately.

Table 3 Weighted average Precision, Recall, F1-measure, and Support related to multi-label classification

Model	W. A. Precision	W. A. Recall	W. A. F1measure	W. A. Support
MLP+T	0.89	0.79	0.84	533
MLP+U	0.84	0.77	0.79	530
MNB+B	0.74	0.87	0.79	543
SVM+U	0.78	0.75	0.77	551
MLP+B	0.82	0.71	0.76	555
RF+U	0.86	0.7	0.76	546
GNB+Tri	0.73	0.78	0.75	579
GNB+T	0.86	0.67	0.74	541
SVM+T	0.91	0.68	0.74	559
GNB+B	0.77	0.71	0.73	575
MNB+Tri	0.66	0.83	0.73	539
RF+T	0.83	0.64	0.69	525
DT+T	0.67	0.68	0.67	540
SVM+B	0.71	0.64	0.67	563
KNN+U	0.75	0.62	0.66	552
MNB+U	0.75	0.6	0.66	546
RF+B	0.75	0.64	0.66	550
DT+U	0.68	0.7	0.65	561
GNB+U	0.75	0.6	0.65	546
LR+T	0.76	0.45	0.6	549
MLP+Tri	0.66	0.61	0.6	564
KNN+T	0.82	0.53	0.59	544
DT+B	0.6	0.58	0.58	560
RF+Tri	0.69	0.57	0.55	571
MNB+T	0.79	0.52	0.52	568
SVM+Tri	0.72	0.43	0.51	558
LR+U	0.75	0.46	0.5	552
KNN+B	0.69	0.45	0.48	581
KNN+Tri	0.53	0.51	0.41	536
DT+Tri	0.71	0.35	0.35	551
LR+B	0.4	0.32	0.27	558
LR+Tri	0.21	0.3	0.25	568

History	0
Culture	0
Economic	1
Health	0
Politic	1
Scientific	0
Sport	0
Technology	0
News	1



اتمر: د مزار- کابل- پېښور د پټلۍ لار
نقشه لاسلیک شوه

۱ ساعت وړاندې

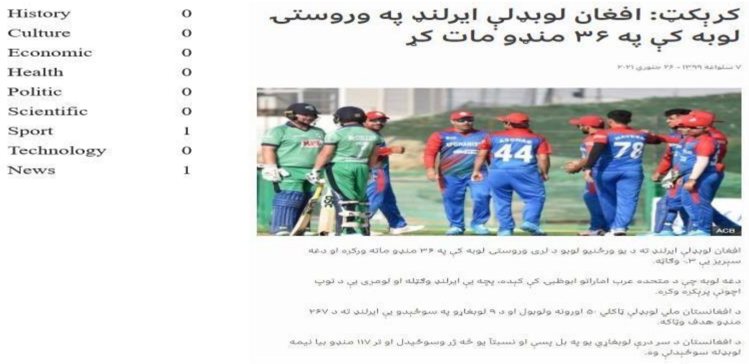
د افغانستان د بهرنیو چارو وزیر حنیف اتمر وایي، د دغه هېواد، پاکستان او ازبکستان ترمنځ د
پټلۍ جوړولو لار نقشه پرون (سې شنبه، دلوه/سلواغه ۱۴) اړوند چارواکو لاسلیک کړه.

حنیف اتمر چې درې ورځېزه غونډه کې د ګډون لپاره د ازبکستان پلازمېنې تاشکند ته تللی خپل ټوېټر
کې دا خبره لیکلې ده.

د نوموړي په اند، د پټلۍ دا ګډه پروژه به د سیمه ییزو اړیکو، سوداګرۍ او ټرانزیټ د ودې لپاره
مشترک ټکی شي او د "افغانستان له لارې د منځنۍ او سوېلي اسیا د نښلولو" ګډه هیله به ترسره
کړي.

ښاغلی اتمر د دې ترون د لاسلیک لپاره د مراسمو برابرولو له امله د ازبکستان د مرستیال لومړي
وزیر عمر زاقوف مننه کړې ده.

Figure 5 Pashto news article example 1
(https://www.bbc.com/pashto)



History	0
Culture	0
Economic	0
Health	0
Politic	0
Scientific	0
Sport	1
Technology	0
News	1

کرېکټ: افغان لوبډلې ایرلنډ په وروستۍ
لوبه کې په ۳۶ منډو مات کړ

۷ سلواغه ۱۳۹۹ - ۲۶ جنوري ۲۰۲۱

افغان لوبډلې ایرلنډ ته د یو ورځنیو لوبو د لړۍ وروستۍ لوبه کې په ۳۶ منډو ماته ورکړه او دغه
سیریز یې ۳-۰ وګاټه.

دغه لوبه چې د متحده عرب اماراتو ابوظبۍ کې کېده، پچه یې ایرلنډ وګټله او لومړی یې د توپ
اچونې پرېکړه وکړه.

د افغانستان ملي لوبډلې ټاکلي ۵۰ اوورونه ولوبول او د ۹ لوبغاړو په سوځېدو یې ایرلنډ ته د ۲۶۷
منډو هدف وټاکه.

د افغانستان د سر دري لوبغاړي یو په بل پسې او نسبتاً یو څه ژر وسوځېدل او تر ۱۱۷ منډو بیا نیمه
لوبډله سوځېدلې وه.

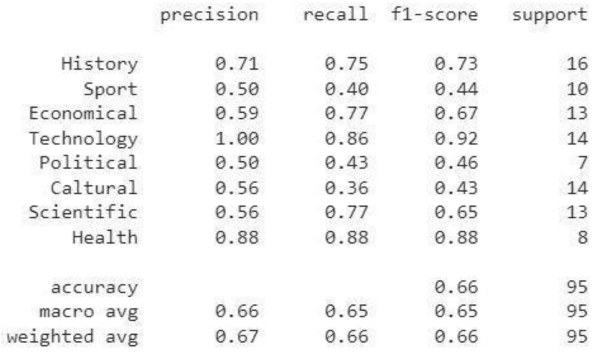
Figure 6 Pashto news article example 2
(https://www.bbc.com/pashto)

7.3. Evaluation of the Third Model

We divided the dataset into 80% and 20% portions for train and test sets, respectively in multilabel classifier. By using DistilBERT-base-multilingual-case the obtained accuracy for document multiclass single label classification is 66.31% (table 5). This model achieved 0.68 hamming score and 0.10 hamming loss in Pashto multiclass multilabel classification task (table 4).

The main reason behind low accuracies of pre-trained language models is that we used multilingual base models. It is trained for more than hundred languages. However, Pashto is not in that list. It is difficult for the model to distinguish and recognize Pashto alphabet characters and morphemes. Thus, these models demonstrate cheap performance in Pashto NLP tasks. Therefore, the requirement of a Pashto tokenizer is perceived.

On the other hand, existence of several similar words in some distinct categories results on misclassification of the labels as illustrated in figure 7.



```
              precision    recall  f1-score   support

     History       0.71      0.75      0.73        16
       Sport       0.50      0.40      0.44        10
  Economical       0.59      0.77      0.67        13
  Technology       1.00      0.86      0.92        14
   Political       0.50      0.43      0.46         7
    Caltural       0.56      0.36      0.43        14
  Scientific       0.56      0.77      0.65        13
      Health       0.88      0.88      0.88         8

    accuracy                           0.66        95
   macro avg       0.66      0.65      0.65        95
weighted avg       0.67      0.66      0.66        95
```

Figure 7 Multiclass classification report using DistilBERT

Table 4 Experimental results of the multilabel classification 3rd group models

Model	Hamming score	Hamming loss
M- DistilBERT	0.68	0.10
M-BERT	0.58	0.14

8. CONCLUSION

This paper is one of the first state-of-the-art research in Pashto literature text classification analysis. It built the first Pashto documents corpus in two versions one for single and the other for multi-label classification purposes. It also made a lexicon list of Pashto words and developed a multiple classification framework to categorize

Pashto documents. This study obtained high accuracy with some classifiers. The highest accuracy achieved by implementing MLP with TFIDF methods in both contexts. The future task is to develop a Pashto tokenizer based on BERT models. Additionally, we will expand our dataset and add a lemmatization task. Moreover, we will observe more state-of-the-art techniques.

REFERENCES

[1] Al-Salemi, Bassam, M. Ayob, G. Kendall, and S. Noah. 2018. “RTAnews: A Benchmark for Multi-Label Arabic Text Categorization.”

[2] Baygin, M. (2018). Classification of Text Documents based on Naive Bayes using N-Gram Features. 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 1–5. https://doi.org/10.1109/IDAP.2018.8620853

[3] Dai, Mian, and Chao-Lin Liu. 2020. “Multi-Label Classification of Chinese Judicial Documents Based on BERT.” Pp. 1866–67 in 2020 IEEE International Conference on Big Data (Big Data). Atlanta, GA, USA: IEEE.

[4] Elnagar, Ridhwan, and Omar E. 2020. "Arabic text classification using deep learning models." Information Processing & Management 57(1):102121.

[5] Farahani, Mehrdad, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. “ParsBERT: Transformer-Based Model for Persian Language Understanding.” Neural Processing Letters 53(6):3831–47. doi: 10.1007/s11063-021-10528-4.

[6] Ghasemi, S., & Jadidinejad, A. H. (2018). Persian text classification via character-level convolutional neural networks. 2018 8th Conference of AI & Robotics and 10th RoboCup Iranopen International Symposium (IRANOPEN), 1–6.
https://doi.org/10.1109/RIOS.2018.8406623

[7] Gutiérrez, B. J., Zeng, J., Zhang, D., Zhang, P., & Su, Y. (2020). Document Classification for COVID-19 Literature. ArXiv:2006.13816 [Cs]. http://arxiv.org/abs/2006.13816

[8] Gutiérrez, Bernal Jiménez, Juncheng Zeng, Dongdong Zhang, Ping Zhang, and Yu Su. 2020. “Document Classification for COVID-19 Literature.” ArXiv:2006.13816 [Cs].

[9] Kamal, U., Siddiqi, I., Afzal, H., & Rahman, A. U. (2016). Pashto Sentiment Analysis Using Lexical Features. Proceedings of the Mediterranean Conference on Pattern Recognition and Artificial Intelligence
- MedPRAI-2016, 121–124. https://doi.org/10.1145/3038884.3038904

[10] Mohtashami, E. and Bazrafkan, M., 2014. The Classification of Persian Texts with Statistical Approach and Extracting Keywords and Admissible Dataset. International Journal of Computer Applications 101, 5, 18–20. https://doi.org/10.5120/17683-8541.

[11] Pal, K., & Patel, Biraj. V. (2020). Automatic Multiclass Document Classification of Hindi Poems using Machine Learning Techniques. 2020 International Conference for Emerging Technology (INCET), 1–5. https://doi.org/10.1109/INCET49848.2020.9154001

[12] Pervez, A. M., Jiangbin, Z., Naqvi, I. R., Abdelmajeed, M., Mehmood, A., & Sadiq, M. T. (2020). Document-Level Text Classification Using Single-Layer Multisize Filters Convolutional Neural Network. IEEE Access, 8, 42689– 42707. https://doi.org/10.1109/ACCESS.2020.2976744

[13] Qadi, Leen Al, Hozayfa El Rifai, Safa Obaid, and Ashraf Elnagar. 2019. “Arabic Text Classification of News Articles Using Classical Supervised Classifiers.” 2019 2nd International Conference on New Trends in Computing Sciences (ICTCS). doi: 10.1109/ICTCS.2019.8923073.

[14] Tegey, Habibullah, and Barbara Robson. 1996. A Reference Grammar of Pashto. Washington, D.C.: Distributed by ERIC Clearinghouse.

[15] Tokgoz, Meltem, Fatmanur Turhan, Necva Bolucu, and Burcu Can. 2021. “Tuning Language Representation Models for Classification of Turkish News.” Pp. 402–407 in 2021 International Symposium on Electrical, Electronics and Information Engineering, ISEEIE 2021. New York, NY, USA: Association for Computing Machinery.

[16] Zahoor, Shizza, Saeeda Naz, Naila Habib Khan, and Muhammad I. Razzak. 2020. “Deep Optical Character Recognition: A Case of Pashto Language.” Journal of Electronic Imaging 29(2):023002. doi: 10.1117/1.JEI.29.2.023002.