Pakistan Journal of Engineering and Technology, PakJET
Multidisciplinary & Peer Reviewed
Volume: 4, Number: 4, Pages: 79- 85, Year: 2021

Pakistani News Classification Based on Headlines

Raza Mukhtar, Muhammad Javed Iqbal, and Zaid Bin Faheem

Department of Computer Science, University of Engineering and Technology, Taxila, Pakistan
*Corresponding author: Raza Mukhtar (e-mail: raza.mukhtar@students.uettaxila.edu.pk)

Abstract- **In information dissemination, media plays a vital role, a huge amount of information is spread quickly through social media and news platforms. The online news outlet and frequency of the huge number of headlines creation demand automatic classification tools. In this overwhelming domain, different countries explored classification algorithms for their news. However, very limited research is done on Pakistani news headlines in terms of classification. There is also a lack of a Pakistani news headlines benchmark dataset. This research is intended to find a suitable model to classify Pakistani news headlines automatically. Moreover, in this study, we designed a benchmark NCE-2D (News Classification and Emotion Detection Dataset) of 2429 Pakistani news headlines extracted from different news websites using ParseHub, which includes five types of categories. First, to minimize the noise from the dataset is undergoing some pre-processing steps including punctuation, stopwords, null entries, and duplication removal. In the next step, extract the feature by two different methods Count-Vectorizer and TF-IDF Vectorizer, and then apply a set of 10 different learning algorithms for both types of features extracted. The best performing algorithms for Pakistani news headlines out of the set of implemented algorithms by both combinations included Support Vector Machine (SVM) when implemented over the TF-IDF features and Multi-Layer Perception (MLP) when implemented over the CV features. Based on the results, both these algorithms are compared to choose the most suitable from them. The combination of SVM is investigated most appropriate classifier with 82.16% accuracy than 81.75% accuracy of MLP.**

Index Terms-- SVM, TF-IDF, News Classification, CV, NCE-2D.

I. INTRODUCTION

With the advancement in time, the value of social media and websites for information gaining is enhanced, like the smartphones, laptops, tabs and many more technologies come into the arena for time-saving and that can be used everywhere with the facility of internet. These advancements of technology play a very important role in services like information gaining, communication, billing, shopping, and transportation. Before these technologies, we use these services manually by ourselves or through colleagues, friends, or family members. Similarly, we also take help from friends and colleagues to explore the hot news of the day, but it is difficult to analyze or find the opinion of others.

In recent days, almost every news channel owns a website, different social media accounts to spread the news everywhere to their viewers. Websites and social media are an arena for exchanging information. There is a huge amount of information on the websites of every news channel, because of that huge amount of data it's difficult to analyze the data manually, filtering the related information from the website.

Currently, there are lots of researchers published different techniques and ways to analyze the data related to news like headlines classification, news article classification, and sentiment analysis by using Machine Learning, Natural Language Processing, and Artificial Intelligence-based tools as I reviewed, Wongso et. al., (2017) discussed the news article classification for the Indonesian news industry using Multi Naïve Bayes classifier and designed the dataset by extracting data from a local Indonesian news website [1]. For the implementation of different methods, a benchmark dataset of related work is needed. So, some of the researchers are designing the informational/featured based trademark datasets by gathering news articles or news headlines base data for their methods, as reviewed William, et al., (2020) designed a labeled dataset for Indonesian news and extract data from twelve different Indonesian news sites of almost 15,000 articles named CLICK-ID [2], and some are using the predesigned dataset to apply new methods or improve the accuracies of existed methods by using different feature selection and Preprocessing techniques.

As we know, the human being can analyze and get the general idea from the text more accurately as the human mind can understand the natural language, but it's difficult to analyze the huge data and get the piece of related information from it manually. To minimize that difficulty researchers introduced various methods that are implemented for a different type of analysis like news classification, detecting business events, fake news detection, sentiment analysis, etc. Qian, Yu, et al. (2019) introduced a method named "clustering annotation classification" a three-step process to detect the business events from massive online news headlines and leads [3].

Our research has proposed a technique to analyze the Pakistani news headlines, which is required for a few reasons. Firstly, very limited research is done on Pakistani news. Secondly, there is also a lack of Pakistani news headlines benchmark dataset, as Kareem, et al., (2019), proposed a model especially for fake news detection for Pakistani news and he designed the dataset of 344 news articles by scraping one of a famous news website [4]. Finally, our designed benchmark dataset and analysis technique can help the media channels to classify headlines on their websites automatically in their respective category, which helps viewers to find out the specific news easily.

II. RELATED WORK

Conducting critical research or making everyday decisions by us often look for other people's opinions. We consult different political discussion forums when casting a political vote, read consumer reports when buying appliances, ask friends to recommend a restaurant for the evening. And now the Internet

has made it possible to out the opinions of millions of people on everything from the latest gadgets to political philosophies. Social media now commands over 22% of the world's total time spent online with 65% of adult internet users using a social networking site. The Internet is increasingly both a forum for discussion and a source of information for a growing number of people. As a response to the growing availability of informal, opinionated texts like blog posts and product review websites, sentiment analysis expands the traditional fact-based text analysis to enable an opinion-oriented information system.

Wongso et. al., (2017) proposed the method of news article classification for the Indonesian language. Extracted data from a local news site, applied many classification techniques, and found the best one as a multi naïve Bayes classifier with a 98.4% precision and recall score [1]. This is a data article in which William, et al., (2020) designed a labeled dataset of Indonesian news from twelve different news sites with almost 15,000 articles named CLICK-ID. Then validate the performance with CNN and Bi-LSTM models and get efficient accuracy values [2]. There is massive data about everything available online, Qian, Yu, et al. (2019) investigated. In this work, the author proposed a method named "clustering annotation classification" a three-step process to detect the business events from online news headlines and leads. First more related data is extracted and processed in a different type of business events cluster then annotate each cluster and utilize the extracted data for classification of potential business events from online business events news headlines and leads. They used WR clustering, Linear Regression, Support vector machine, Naïve Bayes, and in results author found WR clustering performed best with a 64.08% avg. precision, 69.70% avg. recall and 63.62% avg. F-value [3]. Kareem, et al., (2019), proposed a method, especially for fake news detection in Pakistani news. They extracted the data from some popular news sites in Pakistan. Applied different algorithms for it, as a result, they found KNN with 70% of accuracy [4].

Silva, et al. (2020) are defined the problem of fake news detection is available only in the English language, they proposed a technique to detect fake news in the Portuguese language. Also presented labeled dataset for the training and result evaluation. They extracted features by BoW, Word2vec, and FastText and applied multiple classifiers for fake news detection. They found Logistic Regression with BoW is the best method with 97.1% F measures [5]. Mulahuwaish, Aos, et al. (2020) focused on reducing the complexity of time and space of big data processing, which is obtained by web crawling. They considered data mining of just specific and featured data and developed a technique for the web that extracts the news data and is categorized in its specified category out of four defined categories. Also, compare the four different classifiers of ML including SVM, DT, KNN, and LSTM based on accuracy and ROC curve. As result, they found KNN is worst with 88.72% accuracy and SVM is best with 95.04% accuracy [6]. Bahad, et al, (2019) investigated a method for fake news detection, which is a challenge nowadays because of spreading a lot of fake news daily. Proposed the Bi-directional method of LSTM recurrent neural network for fake news classification based on the local dataset. This method applied different techniques on two different data sets and the Bi-directional method of the LSTM recurrent neural network got maximum accuracy against DS1 (real_or_fake) and DS2 (Fake News detection) of 91.08% and 98.75% respectively [7]. Gadek, et al., (2020) published a method for biased fake news detectors which is demanding domain in the recent era. They investigated the TC-CNN technique that detects and classify the news in their respective class fake or real. Additionally, this method is evaluating results automatically and classified the news emotions also. Got best results as 91.8% on fake news detection and 68.6% on emotion classification [8]. Rambaccussing et al., (2020) proposed the method of inflation, unemployment, and output forecasting from the newspaper of the US-based on economic policy-related content printed in it. They used Tf-Idf for feature extraction and the LSVM classifier provides an accuracy of 76.55%. Forecast the results based on graphical analysis [9]. Samuels, et al., (2020) proposed a new sentiment analysis technique on news headlines. They extracted features using Stanford CoreNLP, Tf-Idf, and count vectorizer and evaluate the different models on the base of the smallest and largest dataset. They concluded that linear SVM is performed better with 89.49% accuracy 7 for the smallest and SGD is better than other models for a large dataset with 88.13% accuracy [10].

Samuels, et al., (2020) proposed a method of sentiment analysis for news. Because of huge data in form of blogs, comments, news sharing, etc. is very difficult to process manually. Thus, they proposed a lexicon-based technique that processed sentiment analysis of news data automatically and efficiently to detect their emotion bad, good, or neutral. They evaluate models proposed named WordNet and SentiWordNet over the BBC news dataset [11]. Hui, et al., (2017) analyzed the key text out of original data which is beneficial for emotion detection in news headlines and classification. Then they evaluated the model over based on sentiment and polarity scores individually, they concluded that results are better while using polarity scores for emotion classification with 83.7% of accuracy. However, got an F score of 42.2% when using emotions for classification [12].

Chavan, et al., (2019) presented the SentiWordNet based semantic-oriented approach with the machine learning algorithm for real-time understanding of the USA sentiment about India using the USA newspaper headlines related to India. They used multiple ML techniques and found the best one as SVM with 76.5% accuracy [13]. J SreeDevi, et al., (2020) proposed a classifier for news article classification in their respective category by applying the CNN technique of ML and comparing it with other ones based on training time, prediction time, and their accuracies. They compare CNN with KNN, SVM, and NB algorithms. Found SVM is the best with 90.8% accuracy and taking minimum prediction time, KNN is the best in training with minimum training time than others [14]. Aslam, Faheem, et al., (2020) analyzed the impact of COVID-19 over the news headlines collected over the whole world, which create a crucial mental wellbeing environment globally. They extracted all the headlines with the word coronavirus and analyzed their emotions. They analyzed eight emotions in results using National Research Council Canada Word Emotion Lexicon a high polarity rate of 52% which evoked negative emotion, 30% evoked positive emotions and 18% are neutral. Fear, anger, and sadness are the main emotions in negative polarity [15].

III. METHODOLOGY

The proposed methodology is the detailed visualization of the work that we have done. Figure 1 briefly explained the flow of the work. As we can see the first step is to collect the data from different authentic websites. Then designed the dataset of headlines with the necessary information and features that we need for the classification and opinion mining.

Pre-processing is the next phase of our methodology, to clean and remove the noise from the dataset, which makes the data more efficient to analyze. Once the dataset is cleaned according to the requirements, then converted the data into the numerical or vector form by using the two different techniques including CV and Tf-Idf, then the features of both types are divided into the training, which is used for learning the algorithm and testing data is used for classification testing and then evaluate the results in the end and compare the algorithms performed best based on features extracted from both the techniques.

A. Dataset Designing

When we collect all the data from different websites including BBC Pakistan and allpakistaninews.com by the method of data extraction, according to the required features that are used for the classification process. The dataset designed is named as News Classification and Emotion Detection Dataset (NCE-2D). As discussed in Table I, there are five categories of this designed dataset.

Table I. Dataset Details

Categories	**No. of News**
Technology	494
Financial	398
Investigation	188
Politics	571
Criminal	778
Total	**2429**

Classes of the designed dataset are:

i. Headlines
ii. Headline_url
iii. Short Description
iv. Day & Date
v. Category

The dataset is available at:
https://www.kaggle.com/razamukhtar007/news-classification-and-emotion-detection-dataset.

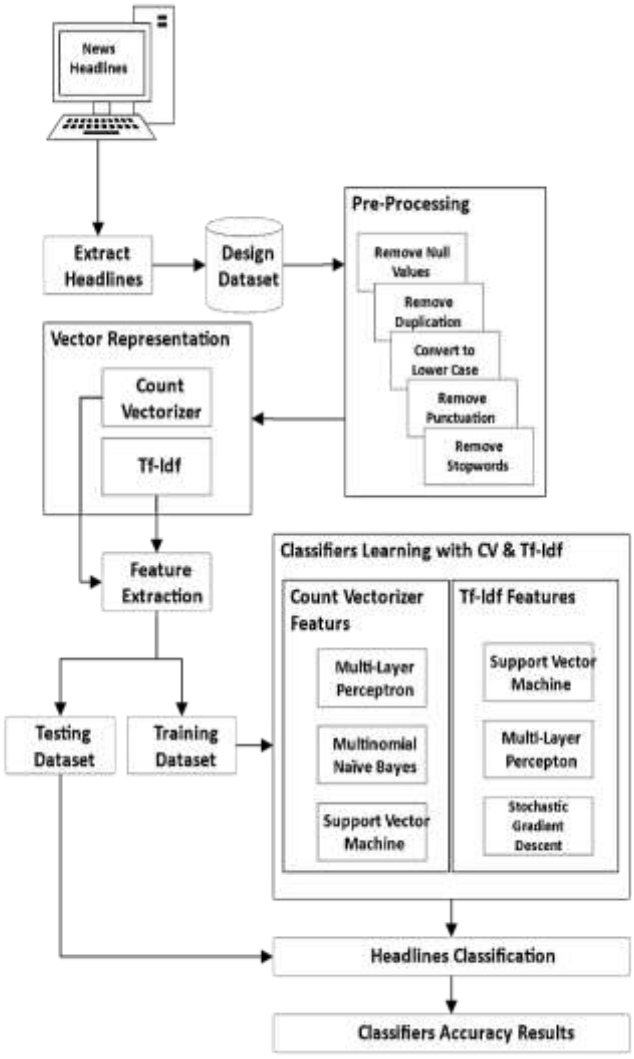

FIGURE 1. Proposed Methodology of Work

B. Data collection

Data collection is a very important part that played a vital role in any research. As needed for work (news classification), required text type data in supervised form like in the form of CSV or excel. After that, data is cleaned by some pre-processing steps.

C. Web scraping

Web scraping is a step in Natural Language Processing. Which collected data from online platforms like different websites including the BBC official website, Channel's official website, and some other authentic sources. In this work, data is collected by the ParseHub tool. Because of some key figures of it better than other tools.

D. ParseHub

ParseHub tool is user-friendly, extracts the data more efficiently, and is freely available to extract just to the point data of 200 pages in one time from a web in supervised form automatically according to the format we set for the first page.

E. Word cloud

Word cloud is a term that defines the collected data based on the text that occurred in the data with high frequency, text with huge size in the cloud is the most repeated word of the dataset. The Word cloud of NCE-2D is shown in Fig. 2.

FIGURE 2. Word cloud of Designed Dataset (NCE-2D)

F. Pre-Processing

Pre-Processing is a process of Natural Language Processing to clean the data, is played a vital role to improve the performance of text classification via learning algorithms, it includes some steps to clean the dataset to make it more efficient for feature selection and classification. Some of these steps are included, convert the text to lower case, remove the stopwords, remove the punctuation, tokenization, stemming, etc. The flow of pre-processing is shown in Fig. 3.

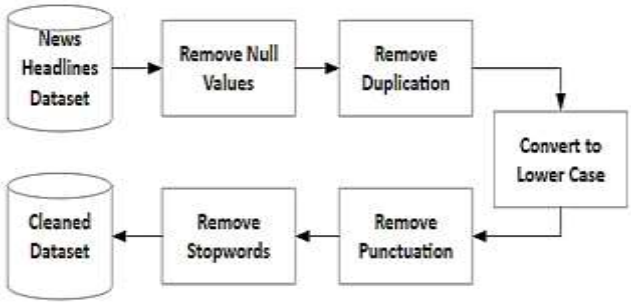

FIGURE 3. Pre-Processing Flow for Data Cleaning

G. Remove Null Values

In this step of preprocessing, data is analyzed to remove the null values in the dataset. To make it more convenient for the learning of classification algorithms. Several algorithms are not accepting the null value, which can cause an error during the implementation of learning and testing algorithms.

H. Remove Duplication

To make all the entries of the dataset are unique, remove all duplicated entries from the dataset. If most entries are duplicated, it causes biasedness of the learning model.

I. Convert Text to Lowercase

In this step, the desired text is converted to lowercase, which is more understandable for a machine. Converted only these features that were required to use for the training and testing purpose.

J. Remove Stopwords

Stopwords removal is the key step to make the sense of sentence clearer to understand, stopwords included the words like I, his, your, these, which, have, me, she, am, etc. Removal of stopwords helps to improve the meaning of sentences, only meaningful words are left behind in text.

K. Remove Punctuation

Punctuation is one type of noise; removal of punctuation symbols makes the models faster as models do not process them easily and are difficult to understand.

L. Feature Selection

When the collected data is cleaned, separated the features from the dataset, suitable for models, which will be used for the training and testing of the model for this work. A classifier can only understand the numbers, this needs to convert the textual representation of data into vector representation of the data.

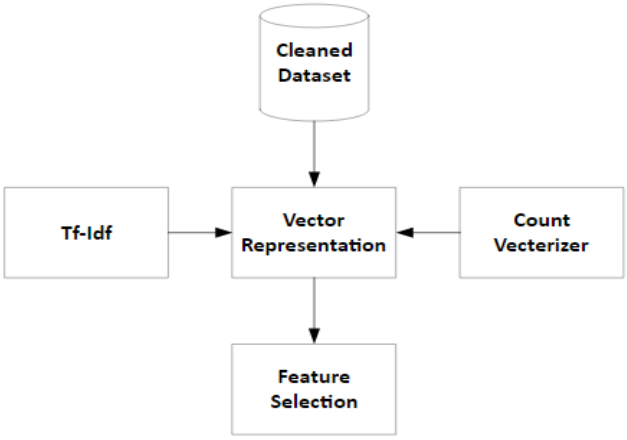

FIGURE 4. Feature Selection Using Tf-Idf and CV

There are many methods to convert text to vector, vectorization using count vectorizer and Tf-Idf vectorizer discussed [5]. We have also used these methods for vector representation of data. Flow is shown in Fig. 4.

M. Count Vectorizer

Count vectorizer is also called term frequency, which determines the frequency-based value of the token. It saves the occurrence of every separate token in the text, the higher the value of the token the higher the occurrence of the token in the text.

N. Tf-Idf

Tf-Idf is one further step for data's vector representation, which is called term frequency and inverse term frequency. It determines the frequency of a token in the text with the specificity of the token into the account. For example, if any token has a higher frequency in all the texts, Tf-Idf assigns a lower value to

this token. The value of the token is higher which occur in a few texts but less in other text.

O. Classification

Several classification algorithms are used for the classification of text. In this work, a set of algorithms is applied with the combination of two different feature extraction techniques. The combination of Tf-Idf with Support Vector Machine and on the other hand Count Vectorizer with Multi-Layer Perceptron is performed best.

P. Support Vector Machine

Support Vector Machine is the binary classifier that divided objects into two classes according to their classes, it determined the object belongs to that specific class or category or not. To find the class of an object a hyperplane is used to map the values of entries of the dataset against their respective class or category. A linear function begins when features are passed through the SVM to determine the boundary of class for dividing the objects into classes. The boundaries that are determined are based on the distance of the object from their class, then the nearest object to any class is added to that class. The object that is closest to its respective class is called the support vector.

Q. Multi-Layer Perceptron

Multi-Layer Perceptron is a neural network's feed-forward supplement classifier for classification. There are three types of layers in MLP, including the input layer, output layer, and hidden layer. MLP mapped the input vector and corresponding output layer nonlinearly, the extracted features are passed through the MLP model. Features are taken as input and classification is performed by the output layer. The real computational engine of MLP is placed in the hidden layer based on an arbitrary number of layers placed between the input and output layer. By using backpropagation learning algorithms of MLP neurons are trained. Mainly, MLP is used for pattern classification, approximation, prediction, and recognition. For the approximation process, more than one hidden layer is required. Neurons' computation at the output and hidden layer is taking place explained as:

$$o(x) = G\big(b(2) + W\ (2) * h(x)\big) \qquad (1)$$

$$h(x) = \ \varphi(x) = s(b(1) + W(1)x) \qquad (2)$$

In the Equation (1) and Equation (2) computations, b(1) and b(2) are biased vectors, W(1) and W(2) are weight matrices, G and s are the activation functions that can select any activation function from several choices of activation functions are available like tanh, sigmoid, etc.

IV. Results and Discussion

Performance metrics are used to measure the performance of algorithms, there are different performance measures, some common (Accuracy, Recall, F1 Score, and Precision), that we used to analyze the performance and founded the best algorithm regarding this work.

We discuss all these measures based on the following actual and prediction-based metrics Table II.

TABLE II. Results Evaluation Metrics

Evaluation Metrics		**Actual**	
		Positive	Negative
Prediction	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Accuracy measure is used to check that how many predictions are made by a classifier are predicted correctly. However, this measure is not much convinced regarding the dataset that is not balanced due to the issues of biased predictions towards the category that is high in frequency and predicts the other categories wrongly.

$$\boldsymbol{Accuracy} = \frac{(TP + TN)}{(TP + FP + FN + TN)} \qquad (3)$$

Where TP is True Positive, TN is True Negative, FN is False Negative, and TP is True Positive. Similarly, ***Recall***, ***Precision,*** and ***F1-Score*** are used to measure the performance of the classifiers. These are much better than that of accuracy to determine the performance of the algorithms when the dataset is not well balanced, and their formulas are given below.

$$\boldsymbol{Precision} = \frac{(TP)}{(TP + FP)} \qquad (4)$$

$$\boldsymbol{Recall} = \frac{TP}{(TP + FN)} \qquad (5)$$

$$\boldsymbol{F1\ Score} = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (6)$$

We discuss all these measures by confusion metric for all best performing algorithms and measure their performance.

The results of these classification algorithms with high accuracy that we implemented with features of Count Vectorizer and Tf-Idf techniques to find the best suitable algorithm. So, now here we discuss the results of the best-performed algorithms in both combinations named classification with count vectorizer and classification with Tf-Idf in proposed methodology shown in Fig. 1 and to measure the performance using evaluation metrics in Table II.

A. Classification with Tf-Idf

Classification with Tf-Idf is the first combination of a set of classifiers, that used features extracted from the Tf-Idf technique. There are ten different learning algorithms are included in this set, which is mostly used in different research works shown in Table 3 and performed better in their standards of news category. Here we implement this set of algorithms for Pakistani news headlines to find the most suitable algorithm. Here in Table III, we discussed each model via four different performance measures including Accuracy, F1-Score, Precision, and Recall,

which are used worldwide to compare and measure the performance of the algorithm based on discussed measures.

TABLE III. Classifiers' Measures When Using Tf-Idf Features

Models	**Accuracy**	**F1-Score**	**Precision**	**Recall**
Support Vector Machine	0.82	0.81	0.82	0.82
Stochastic Gradient Descent	0.81	0.81	0.81	0.81
Multi-Layer Perceptron	0.80	0.80	0.81	0.80
Extra Tree Classifier	0.76	0.76	0.78	0.76
Logistic Regression	0.75	0.73	0.77	0.75
Random Forest	0.74	0.75	0.76	0.74
Multinomial Naïve Bayes	0.73	0.70	0.77	0.73
K-Nearest Neighbour	0.72	0.72	0.74	0.72
Decision Tree	0.70	0.71	0.72	0.70
Bagging Classifier	0.70	0.71	0.74	0.70

Table III discussed the measures using the formulas of accuracy by (3). Similarly, for F1-score, Precision, and Recall used (4), (5), and (6) respectively.

Discussed the classifiers based on their measures, can see SVM is the best-performed algorithm concerning the accuracy of when implemented over the Tf-Idf features for classification.

Based on other measures like F1-score, precision, and recall, SVM is also the best-performed algorithm with 81.85%, 82.10%, and 82.17% respectively.

Classification report of SVM measures according to the categories individually, when using the Tf-Idf features:

	Precision	Recall	F1-score	Support
0	0.87	0.89	0.88	234
1	0.81	0.64	0.71	124
2	0.76	0.64	0.69	55
3	0.81	0.88	0.85	185
4	0.78	0.87	0.82	131
Accuracy			**0.82**	**729**

Where index numbers are encoded labels of categories mentioned in the classification report where 0,1,2,3,4 are encoded labels of Criminal, Politics, Investigation, Financial, Technology respectively.

B. Classification with CV

Similarly, as the first combination, in this combination of algorithms with count vectorizer features. The whole set of algorithms and their measures via four different evaluation measures including accuracy, f1-score, precision, and recall discussed in Table IV.

TABLE IV. Classifiers' Measures When Using CV Features

Models	**Accuracy**	**F1-Score**	**Precision**	**Recall**
Multi-Layer Perceptron	0.81	0.81	0.81	0.81
Logistic Regression	0.80	0.81	0.81	0.80
Multinomial Naïve Bayes	0.80	0.80	0.80	0.80
Stochastic Gradient Descent	0.80	0.79	0.78	0.80
Support Vector Machine	0.79	0.79	0.80	0.81
Extra Tree Classifier	0.75	0.75	0.76	0.75
Bagging Classifier	0.73	0.74	0.75	0.73
Decision Tree	0.72	0.72	0.74	0.73
Random Forest	0.72	0.72	0.74	0.72
K-Nearest Neighbour	0.56	0.54	0.73	0.56

Table IV discussed the measures using the formulas of accuracy by (3). Similarly, for F1-score, Precision, and Recall used (4), (5), and (6) respectively.

Discussed the classifiers based on their measures, can see MLP is the best-performed algorithm with the accuracy of 81.89% when we use CV features for classification.

Based on other measures like F1-score, precision, and recall, MLP is also the best-performed algorithm with 81.89%, 81.62%, and 81.89% respectively.

Classification report of MLP measures according to the categories individually, when using the CV features:

	Precision	Recall	F1-score	Support
0	0.85	0.89	0.87	234
1	0.80	0.59	0.68	124
2	0.66	0.67	0.67	55
3	0.81	0.86	0.83	185
4	0.80	0.86	0.83	131
Accuracy			**0.81**	**729**

Where index numbers are encoded labels of categories mentioned in the classification report where 0,1,2,3,4 are encoded labels of Criminal, Politics, Investigation, Financial, Technology respectively.

V. RESULTS COMPARISON

Now, here we discussed in Fig. 5, the comparison of both combinations (Tf-Idf features and CV features) based on the

accuracy to find the best suitable algorithm for Pakistani news headlines classification by comparing both combinations that we have implemented using the features of two different feature extraction techniques. Graphical visualization of comparison is shown in Figure 5.

When both results are compared, the best algorithm for this work investigated is SVM when Tf-Idf features are used with an accuracy of 82.16% than MLP with 81.89% accuracy.

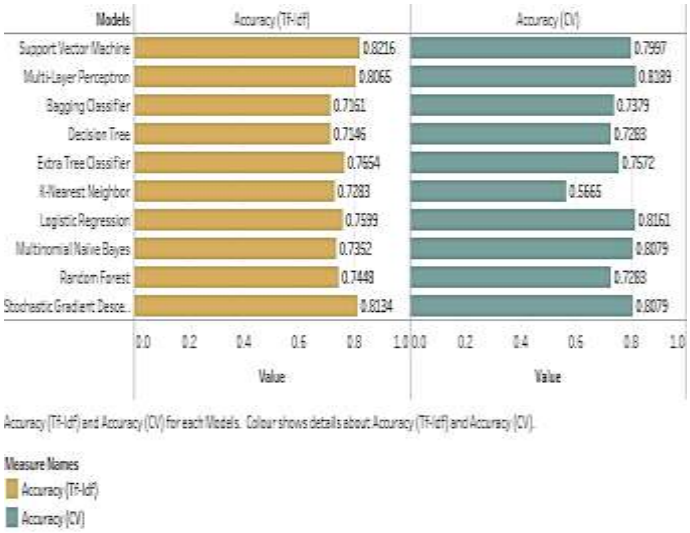

FIGURE 5. Results Comparison of Both Combinations

We also validate the results of our designed dataset by implementing the proposed model best combination with Tf-Idf on an existing dataset named AGs News Classification Dataset mentioned in [16]. In comparison, our model results are performed better for NCE-2D with 82.16% of accuracy than 75.13% of accuracy while performed on the existing dataset.

VI. CONCLUSION AND FUTURE WORK

Keeping in mind the current research domain regarding the classification of news headlines. In this work, we have designed the trademark dataset for Pakistani news headlines and found a suitable classification algorithm.

Our proposed methodology can help the Pakistani news media houses to add the news into their specific category automatically. In the future, we can improve this work by increasing the dataset values. Also can introduce the news authentication process in the model before proceeding further for classification.

ACKNOWLEDGMENT

I owe my appreciation to every one of those individuals of my institution UET Taxila who has made this work conceivable and considering whom my graduate experience has been one that I will value until the end of time.

REFERENCES

[1] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, and O. Rusli, "News article text classification in Indonesian language," *Procedia Computer Science*, vol. 116, pp. 137-143, 2017.

[2] A. William and Y. Sari, "CLICK-ID: A novel dataset for Indonesian clickbait headlines," *Data in brief*, vol. 32, p. 106231, 2020.

[3] Y. Qian, X. Deng, Q. Ye, B. Ma, and H. Yuan, "On detecting business event from the headlines and leads of massive online news articles," *Information Processing & Management*, vol. 56, p. 102086, 2019.

[4] I. Kareem and S. M. Awan, "Pakistani Media Fake News Classification using Machine Learning Classifiers," in *2019 International Conference on Innovative Computing (ICIC)*, 2019, pp. 1-6.

[5] R. M. Silva, R. L. Santos, T. A. Almeida, and T. A. Pardo, "Towards automatically filtering fake news in Portuguese," *Expert Systems with Applications*, vol. 146, p. 113199, 2020.

[6] A. Mulahuwaish, K. Gyorick, K. Z. Ghafoor, H. S. Maghdid, and D. B. Rawat, "Efficient classification model of web news documents using machine learning algorithms for accurate information," *Computers & Security*, vol. 98, p. 102006, 2020.

[7] P. Bahad, P. Saxena, and R. Kamal, "Fake news detection using bi-directional LSTM-recurrent neural network," *Procedia Computer Science*, vol. 165, pp. 74-82, 2019.

[8] G. Gadek and P. Guélorget, "An interpretable model to measure fakeness and emotion in news," *Procedia Computer Science*, vol. 176, pp. 78-87, 2020.

[9] D. Rambaccussing and A. Kwiatkowski, "Forecasting with news sentiment: Evidence with UK newspapers," *International Journal of Forecasting*, vol. 36, pp. 1501-1516, 2020.

[10] C. J. Rameshbhai and J. Paulose, "Opinion mining on newspaper headlines using SVM and NLP," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, pp. 2152-2163, 2019.

[11] A. Samuels and J. Mcgonical, "News Sentiment Analysis," *arXiv preprint arXiv:2007.02238*, 2020.

[12] J. L. O. Hui, G. K. Hoon, and W. M. N. W. Zainon, "Effects of word class and text position in sentiment-based news classification," *Procedia Computer Science*, vol. 124, pp. 77-85, 2017.

[13] S. S. Chavan, "Sentiment Classification of News Headlines on India in the US Newspaper: Semantic Orientation Approach vs Machine Learning," Dublin, National College of Ireland, 2018.

[14] S. D. Mahajan and D. Ingle, "News Classification Using Machine Learning."

[15] F. Aslam, T. M. Awan, J. H. Syed, A. Kashif, and M. Parveen, "Sentiments and emotions evoked by news headlines of coronavirus disease (COVID-19) outbreak," *Humanities and Social Sciences Communications*, vol. 7, pp. 1-9, 2020.

[16] S. D. Mahajan and D. Ingle, "News Classification Using Machine Learning."