Pakistan Journal of Engineering and Technology, PakJET
Multidisciplinary | Peer Reviewed | Open Access
Volume: 03, Number: 03, Pages: 26-30, Year: 2020

Classification of News Articles using Supervised Machine Learning Approach

Muhammad Imran Asad, Muhammad Abubakar siddique, Safdar Hussain, Hafiz Naveed Hassan, and Jam Munawwar Gul

Khwaja Freed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan.
Corresponding author: Muhammad Imran Asad (Email: imranasadkp@gmail.com)

Abstract: **Today the big challenge for News organization to well organize the news and well categorize the news in automatically no need the data entry people to enter and select the category and then based on the category and its sub-category they will be manually selected and enter the details and then after this the analysis will later on used for different aspects. The news is almost every second used in different sources of media in soft and hard. We use the both sources of the Pakistan News in dual languages English and Urdu both and process them and prepare them for machine learning and based on the Machine learning trained data we build a very effective and efficient model that can predict the title category of the news and category of description of the news. We use different machine learning algorithms and different features extraction finally we build the model using the machine learning algorithm with 89% accuracy with logistic regression.**

Index Terms- Classification, Logistic Regression, Machine learning, NEWS, Random Forest.

I. NTRODUCTION

News is the big source of information and the in the day daily activity the NEWS is watching for the purpose of the updated with the current situation of the world. The News media works with different media there are newspapers in which the two type mostly used daily newspapers and weekly newspapers some are local and some are international other media is TV channel the TV channels are working with broadcast media and broadcast the NEWS using the cable TV network this service is provided by Cable TV network service provider and the CTVSP are using the satellites which are enabled to receive the proper signals from the TV channel and provide help to reach at each nod [1]–[3].

The other one source of NEWS information is social media in which the big platform for video sharing the YouTube , YouTube is the mostly frequently used portable media its Application available in the Android Google Play store(Android application store) in all Android devices it's free to download and install and use the YouTube works with Gmail id and it is the product of Google limited liability company (Google LLC) short name is Google [4]. The YouTube platform have many products and features in which the video channel is one of them and due to its robustness, most of the TV channels are developed their own NEWS channels because of the YouTube free service and if the channel subscriber increased and based on the views the Google also pay the channel owner and according to Google policy. In Pakistan there are also every channel have their own YouTube channel. In Pakistan there are almost more than 68 NEWS companies they have their TV channels and as well as YouTube channels and social media pages and social media accounts.

There are 22 Urdu language news channels and 4 are in English language and the remaining are in other local languages. Here is the total information about news channels in Pakistan contribution in different. In Asia as we know the

national language the people have speech the other native languages like Punjabi, Saraiki, Pashto, Sindhi, Balochi shown in Fig. 1. These languages are not the national languages but there is so good even in Pakistan having 4 provinces matched with these 4 languages Punjab speech Punjabi, Sindh speech sandhi and Baluchistan speech Balochi and the last one is the mix-up of different languages which is province name is Sarhad [5].

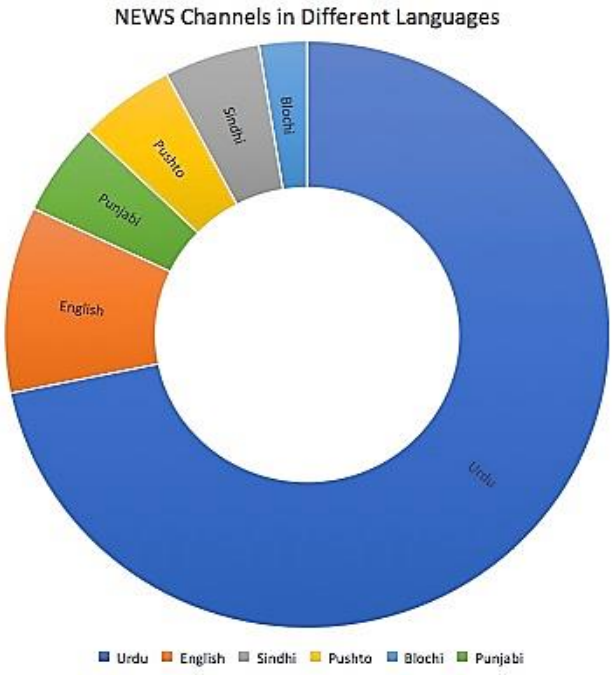

FIGURE 1: Here are the statistics about the NEWS channels in Pakistan with Different languages.

Here we see that the Urdu channels are more than others because the Urdu is the national language in Pakistan. Here are the graphs based on the percentile rating of top 15 channels.

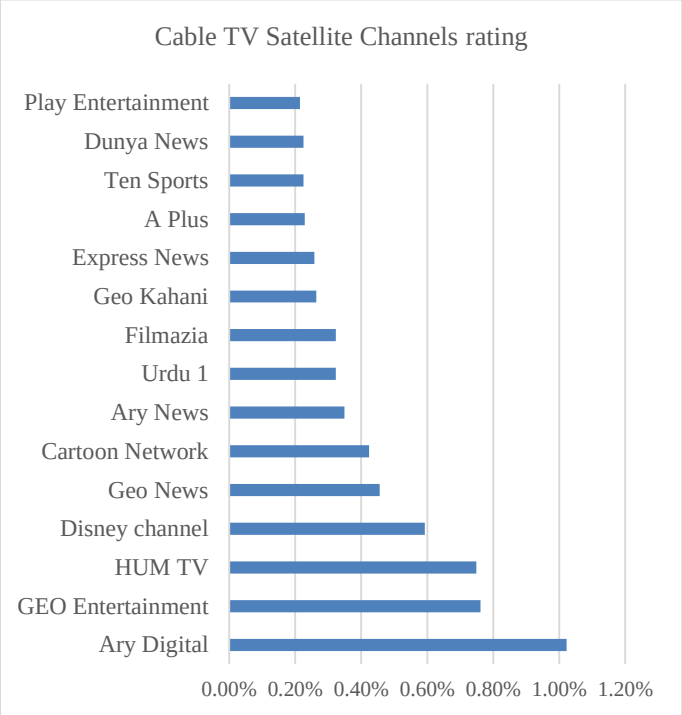

FIGURE 2: Here are the Cable TV well known channels rating by watch.

We also collected the data about the change in the news channels in Pakistan and their growth and also make comparison from 2014 to 2018 this data is 2018 till because its publicly not allowed and accessible [5].We see that the ARY Digital is more good rating in between 2014 to 2018 and the worst rating is A Plus TV channel the one thing we notice that the channels rating is also depends on the team of the channels is the team move the rating of the channels will also move and the channels so dependent on the channels team. New channels are also before the 2014 the GEO was going to very good rating in News and entertainment as well. After the 2014 the GEO News face the competition with different news channels and the News go back and the ARY News coms first today 2020 the new Channels BOL News are also reach the good rating and the data is not accessible for publish use on the government website (PSB)[5]. Here we show the comparison of channels with different languages in Fig. 2 and Fig. 3.

Watching ratings shows that some channels are local channels and some are international channels. Growth of the Channels in Pakistan from 2014 to 2018 by the Department of the statistics Government of Pakistan [5].

We noticed that the channels ratio is changed year by year but the Urdu and English languages channels are almost same so decided to select the two languages News classification due to hugeness of the data the News agencies are need time to input into the system and then after entry the system will managed by the user and the user is takes wrong class or category than the analysis will goes wrong and based on the wrong data we cannot take the right decision. The News is the source of future decisions. News channels are need to manage them automatically based on the Title and description of the class of the News.

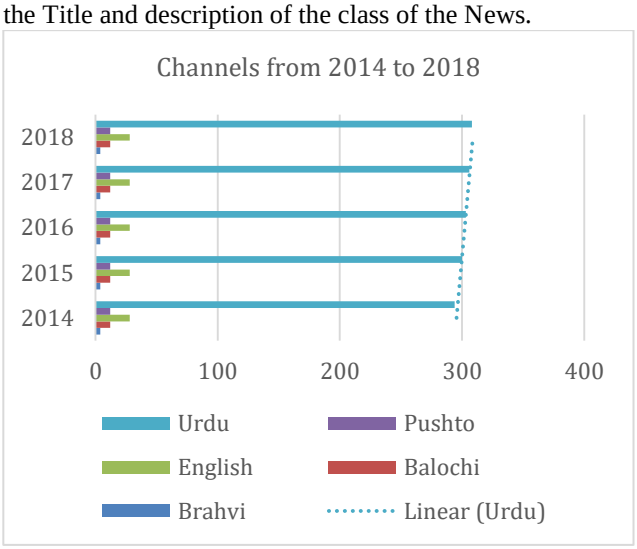

FIGURE 3: The number of channels in Pakistan in quantity with year by year increased and 2018 the figure reached more than 300 channels in the different languages.

We proposed a work here for the automatic classification.

Our Contribution:

1 We proposed dual (Urdu + English) languages News classification support automatically.
2 Our proposed model can predict class based on the Title of the News.
3 Our proposal can predict the category of the class and sub-category based on the News description automatically.

Our proposed work is so effective and efficient method through this we automatically classify and based on the classification the analysis will be productive and efficient.

I. II. LITERATURE REVIEWS

News classification can be done with machine learning algorithms. The algorithms help the user to automate the processes. Study use the machine learning algorithms which are Support Vector Machine, Decision Tree, Naive Bayes, K-nearest Neighbor statistical algorithms and compare the results and find the best one is SVM for news classification. The results show the accuracy 95% [6].

In 2016 researchers proposed a novel approach to classify news data into different groups so that users can read popular news topics. The dataset used by the researcher was BBC news and the 20Newsgroup dataset [7]. They use TF-IDF for feature extraction and to classify news using SVM. Accuracy of the model was desirable that is 97.84%. and 94.93% respectively [6].

In 2017 researcher's compared different machine learning models on four different news datasets which are Reuters, 20Newsgroup, BBC News and BBC Sport dataset. They observed that best accuracy of Linear SVM as compared to other models with accuracy of 86.7%, other models are Naïve Bayes, K-nearest Neighbors, Decision tree and Rocchio algorithm [8] also observed that these algorithms performed well for small datasets. Experiments show that

SVM has more accuracy but Naïve Bayes has better time complexity with less accuracy [7].

In 2018 researchers propose a novel approach to classify Chinese news text classification. The dataset used by them was Fudan University news corpus. They use different machine learning algorithms like KNN, Naïve Bayes and SVM and compare their results. The study shows that SVM with TF-IDF is the best accuracy of 95.7% for the small datasets as compared to Naïve Bayes which is easy to implement. KNN is the best of that dataset which has more overlapped categories [8].
In 2017 researcher comparison of different machine learning algorithms for news classification of Indonesian language, the dataset used by them was crawled by Indonesian news websites. The study shows that TF-IDF with Multinomial Naïve Bayes has more accuracy which is 98.4% compared to other classifiers. The results are more satisfied than the previous study which is 85%[9].
In 2018 researchers used some supervised machine learning methods to classify Azerbaijani news articles and compare different methods before and after the feature selection process[9]. They gather about 130000 news articles. To perform feature selection and preprocessing they used Chi-squared test and LASSO methods. LASSO increased the accuracy of Naïve Bayes from 71.5% to 76.9% and SVM increased by 87.9% to 88.5%. The artificial neural network increased the accuracy from 86.3% to 89.1% using the Chi-Squared test. They experience that we can't guess which preprocessing technique is best because of dataset properties[2].
News can be fake and the social media we know that the the news are mostly fake and this is possible with the help of machine learning and the experiments shows that the model build for Facebook fake deduction with the 74% accuracy on the Naive Bayes algorithm(Liu and Wu 2018).Helps the user to find the fake news and make decision positively timely[10].

II. III. DEPLOYMENTS & EXPERIMENTS

DATA COLLECTION:
We collected data from the different NEWS sites in Pakistan majorly form the Sama TV Dawn News, Daily Time News, web site [5] using the python package *BeautifulSoup* which is free and open source and easy to use another tool we use is instant Scrapper which is the Google Chrome extension that is so effective and useful application for data scraping. some website are not allow the user to download the news and some are too old data allow to download [11].

PREPARING DATA AND EXPERIMENTS:
Our Data is consisting on the three columns one is Title: Which is the Tile or the head line of the News and the next one is Description: The description of the NEWS or details of the NEWS and the last one is classes of the data and with subclasses[12].
The dataset consists of the 600 NEWS and classes dataset. The description of the dataset (see Tab. I).

	COUNT	UNIQUE	TOP	FREQ
title	606	592	Gujarat, Cambodia World Cup 3rd stage ends	2
description	606	595	GUJRAT: The Pakistan Kabaddi Federation has channel	2
category-NEWS	606	4	Health-disease	156

TABLE I: Describes the numbers of titles and the descriptions and what are the number of categories is available in the News Dataset.

Preprocessing helps us to remove the unnecessary data and after cleaning the dataset we use the features engineering. We use the four categories and its sub-category which are , Health in Health we observe the sub-category is what about in health category in our case Health-corona , Health-heart , Health-cancer , and other one category is Sports the subcategory of sports is specially in Pakistan in Kabaddi , Cricket , Hockey now than we use the python languages detector and detect the languages and use the google Translator services to convert the all data into one normalized form which is English and based on the English languages we annotate the dataset and the dataset is annotated by 5 volunteers as shown in Fig. 4 and finally we done our dataset for use [4].

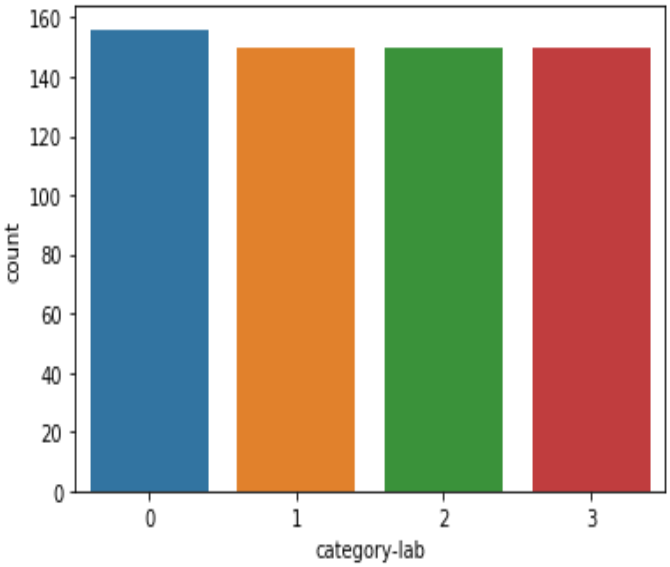

FIGURE 4: Labeled News dataset quantity base information and balancing the number of labels.

After some experiment our predictions are not good because of our classes are not balanced and after the more scrapping and normalization we have balanced classes and then we use the features engineering different techniques which are TFIDF and Word2vec in TFIDF we use the uni-gram, Bigram, uni-bigram [13]. We use the 70% for training and 30% for test and use the Logistic regression and LDA and Random Forest with these we find the best one is logistic regression with F1 score 0.98 and 93% of the classification accuracy[13]–[15].

METHODOLOGY

After the data collection and preprocessing we use the *pycart, sklearn* and for languages detection we use the python package and then use the services of the Google Translator we also use Bing services and translate the NEWS data into English languages. The languages detector is used for Urdu and Roman-Urdu and then uses the translation services. The conversion accuracy is mostly is perfect and both the languages Urdu. Roman-Urdu [16].

Our methodology diagram given in Fig. 5 helps us to understand our working processes and helps the reader to understand the processes and flow of the working methodology. We are use the simple and shortest way to reach our goal.

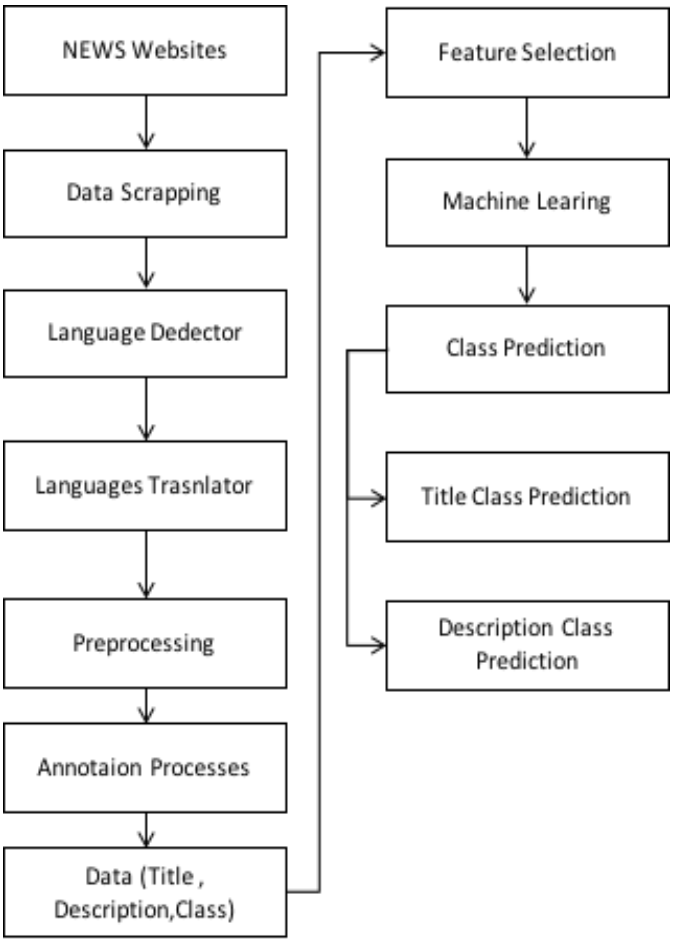

FIGURE 5: Describe the processes model of our methodology and step by step operations we performed.

The proposed model shows the prediction of the title/headlines and description we use this model for experiments and research our goal [7]. The annotation process helps us to build the data for training and based on the title and description we labeled the class for that after the train dataset building the data is ready for feature engineering we use here counter vectorization , TFIDF vectorization , Word2vec and after this the data is fully ready for machine learning we use the two algorithms one is Logistic Regression on these three features and get the results and other is Random Forest is on these there features extraction methods and the effective result is Logistic Regression[9], [10].

III. IV. RESULTS AND DISCUSSION

Remember that we use the hyper parameters to find the best parameters and best results. We use the 10-fold cross validation and we get the uni-bigram is the best feature extraction technique and logger plenty is l2 based on these parameters the description class prediction result is 94% and F1 score is 0.98 shown in Fig. 6. We also work on the title class prediction also here we show the results which are obtained for Random Forest and Logistic regression and its comparison.

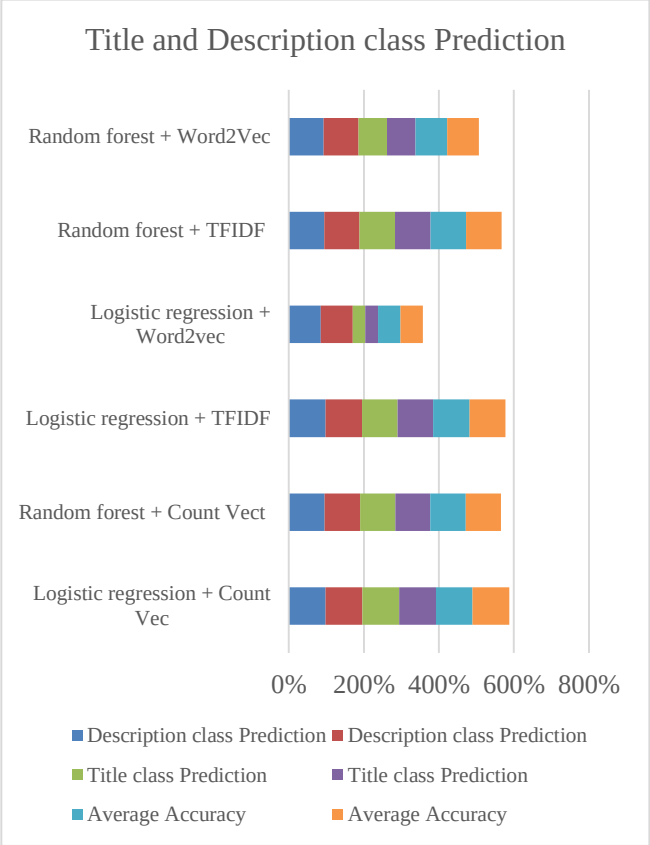

FIGURE 6: Describe the results with different features engineering techniques and performance of the algorithm.

These are the results for description class prediction and accuracy the best on results on the logistic regression which is 98%. We also work on the title class prediction here are the results for title class prediction. We use the title because most of the news channels and news are firstly presented as headlines and the user feel the headline are good for producing the interest for viewer. After experiments shows the results are so productive. We also use the language translator and language detector which is the python package [17] (*LANGDETECT)* is used to detect the language if news in English ignore conversion if in the Urdu / Roman-Urdu than convert the news title and description into English and after this we use this dataset[13], [18-21].

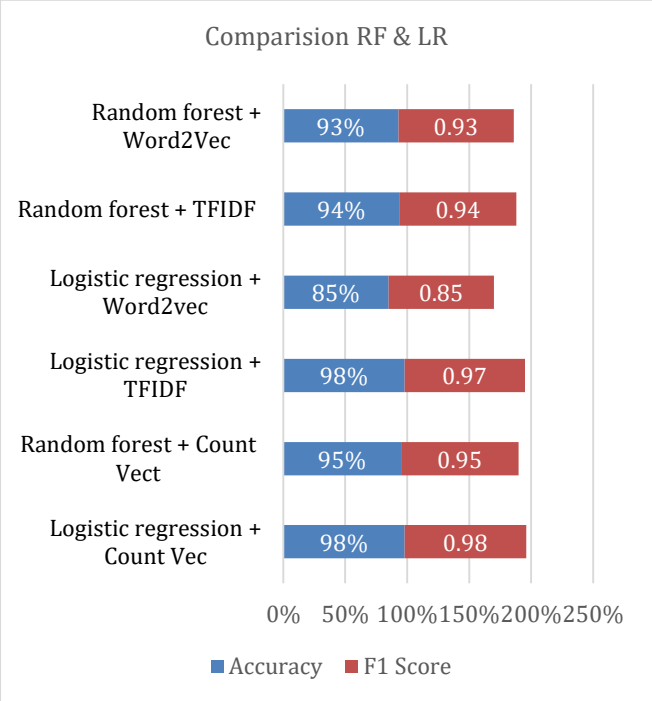

FIGURE 7: The Comparison of the top two models with the comparison of the different features and comparison of F1 and Accuracy performance of the models.

Here we see the results in Fig. 7 about the features having the different results with different features with different algorithms. Our experiments are on the 3 phases Title / Headlines class prediction sent based on the description we can predict the class and the 3rd is use of Title / head line and description to get the impressive results. Here are the results in fig.

Most of the experiments are superior and the working of the model is effective for the classification of the news dataset if in the Title / headline or description our model will help the news companies to automatically classify the news and not to do manual categorization of the data.

IV. V. CONCLUSION

We successfully build the model or NEWS Classification and the effective feature selection method is TFIDF and the best algorithm is logistic regression. Our proposed method is helpful for classification of the NEWS even if we have only Title or even, we have only description or both of them we can automatically classify this in category and sub-category effectively and efficiently. Our model accuracy 98% helps us to confidence us to deploy in the NEWS companies.

REFERENCES

[1] R. Luss and A. d'Aspremont, "Predicting abnormal returns from news using text classification," *Quant. Finance*, vol. 15, no. 6, pp. 999–1012, 2015.

[2] Y. Liu and Y.-F. B. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," 2018.

[3] J. Li, S. Fong, Y. Zhuang, and R. Khoury, "Hierarchical classification in text mining for sentiment analysis of online news," *Soft Comput.*, vol. 20, no. 9, pp. 3411–3420, 2016.

[4] "Google Trans Python Free API." https://pypi.org/project/googletrans/.

[5] "Pakistan Department of Statistics Punjab," 2018 2020. http://www.pbs.gov.pk/content/social-statistics.

[6] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," *Int. J. Res. Mark.*, vol. 36, no. 1, pp. 20–38, 2019.

[7] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani, "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification," in *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, Coimbatore, India, Mar. 2016, pp. 112–116, doi: 10.1109/ICETECH.2016.7569223.

[8] F. Miao, P. Zhang, L. Jin, and H. Wu, "Chinese News Text Classification Based on Machine Learning Algorithm," in *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Hangzhou, Aug. 2018, pp. 48–51, doi: 10.1109/IHMSC.2018.10117.

[9] P. Barberá, A. E. Boydstun, S. Linn, R. McMahon, and J. Nagler, "Automated Text Classification of News Articles: A Practical Guide," *Polit. Anal.*, pp. 1–24, Jun. 2020, doi: 10.1017/pan.2020.8.

[10] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 2017, pp. 900–903.

[11] "BeautisulSoup Python Scrapping Tool," *BeautisulSoup Python Scrapping Tool*. https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[12] S. Al-khateeb and N. Agarwal, "Tools and Methodologies for Data Collection, Analysis, and Visualization," in *Deviance in Social Media and Social Cyber Forensics*, Springer, 2019, pp. 45–65.

[13] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. Merschmann, "Categorizing feature selection methods for multi-label classification," *Artif. Intell. Rev.*, vol. 49, no. 1, Art. no. 1, 2018.

[14] Q. Ke, M. Bennamoun, S. An, F. Boussaid, and F. Sohel, "Human interaction prediction using deep temporal features," in *European Conference on Computer Vision*, 2016, pp. 403–414.

[15] Y. Zhang, D. Gong, X. Sun, and Y. Guo, "A PSO-based multi-objective multi-label feature selection method in classification," *Sci. Rep.*, vol. 7, no. 1, p. 376, 2017.

[16] A. U. R. Khan, M. Khan, and M. B. Khan, "Naïve Multi-label classification of YouTube comments using comparative opinion mining," *Procedia Comput. Sci.*, vol. 82, pp. 57–64, 2016.

[17] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, and M. T. Sadiq, "Automatic Detection of Offensive Language for Urdu and Roman Urdu," *IEEE Access*, vol. 8, pp. 91213–91226, 2020.

[18] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *European Conference on Information Retrieval*, 2018, pp. 141–153.

[19] S. Yasin, K. Ullah, S. Nawaz, M. Rizwan, and Z. Aslam, "Dual Language Sentiment Analysis Model for YouTube Videos Ranking Based on Machine Learning Techniques", Pakistan J Engg & Tech, vol. 3, no. 2, pp. 213-218, Oct. 2020.

[20] M. Akram, M. Siddique, M. Jamil, M. Javaid, and A. Haider, "YouTube Video Recommendation Based Multi-lingual Feedback", Pakistan J Engg & Tech, vol. 3, no. 2, pp. 209-212, Oct. 2020.

[21] S. Nawaz, M. Rizwan, S. Yasin, M. Ahmed, and U. Farooq, "Multi-Class Classification of the YouTube Comments using Machine Learning", Pakistan J Engg & Tech, vol. 3, no. 2, pp. 183-188, Oct. 2020.