ZANCO Journal of Pure and Applied Sciences
The official scientific journal of Salahaddin University-Erbil
https://zancojournals.su.edu.krd/index.php/JPAS

ISSN (print):2218-0230, **ISSN (online):** 2412-3986, **DOI:** **http://dx.doi.org/10.21271/zjpas**

RESEARCH PAPER

Improvement performance by using Machine learning algorithms for fake news detection

Eman Shekhan Hamsheen[1] ,Laith R.Flah[2]

[1]College of Science ,Computer Science and IT Department, Salahaddin University-Erbil, Kurdistan Region, Iraq.
[2]Computer Science Department, Cihan University -Erbil, Kurdistan Region, Iraq.

ABSTRACT:

The prevalence of internet use and the volume of actual-time data created and shared on social media sites and applications have raised the risk of spreading harmful or misunderstanding content, engaging in unlawful activity, abusing others, and disseminating false information. As of today, some studies have been done on fake news recognition in the Kurdish language. For extremely resourced languages like Arabic, English, and other international languages, false news detection is a well-researched research subject. Less resourced languages, however, stay out of attention because there is no labeled fake corpus, no fact-checking website, and no access to NPL tools. This paper illustrates the process of identifying fake news, using two components of the dataset for fake news and actual news. Several classifiers were then applied to the quantity after using identifiers as a highlight of selection. Results of the proposed study demonstrated that Passive-Aggressive Classifier (PAC) outperformed the other classifiers on both datasets the dataset with an accuracy score of 93.0 percent and other classifiers were less in some percentage that show high accuracy as well since it is 90 percent.

KEY WORDS: Fake News Detection, Kurdish Language, Machine learning, Classifiers, Passive-Aggressive.
DOI: http://dx.doi.org/10.21271/ZJPAS.35.2.6
ZJPAS (2023) , 35(2);48-57 .

1.INTRODUCTION :

Researchers from all over the world are very interested in how to categorize news items, posts, and blogs as legitimate or fraudulent. Several findings have been performed to determine the influence of fabricated and false news on the public and how people react to such news. Falsified news, sometimes known as fabricated pos newest, refers to any textual or non-textual content that is phony and created to lead readers to believe untrue information (Khalifa et al., 2019). According to the writers of fabrication of information, there are three primary ways that users of social media networking sites consume news: Text is analyzed by computational linguistics, which systematically and semantically focuses on the origin of text. (multilingually). Since texts make up most posts, extensive effort has been done on their analysis (Abdulrahman et al., 2019).

Multimedia is a single post incorporates various media types. This might consist of sound, video, pictures, and graphics. This is quite beautiful and draws people' attention without making them think about the content. By allowing the author of the post to cross-reference to several sources, hyperlinks help build visitors' trust by attesting to the piece's genesis (Shu et al., 2017). Even embedding of images and cross-referencing to other social media networking sites are done in practice. The various categories of fake news by were defined by research that contacted (Oshikawa et al., 2020) and it will be represented in figure 1 as the visual based, user based, etc.

* **Corresponding Author:**
Eman Shekhan Hamsheen
E-mail: eman.hamsheen@su.edu.krd

Article History:
Received: 21/07/2022
Accepted: 06/10/2022
Published: 20/04 /2023

According on how news represented inside a text, as well as how the unused characters such as (., $, %, *, !, and etc.) as it represented in programming with the phrase of bad characters so in this stage most of the work will be related to the cleaning data and prepare it for training later to evaluate the proposed algorithm a(McEnery and Wilson, 2003).

Visual-based
User-based
Knowledge-based
Style-based
Stance-based

Figure 1: Fake news types

So as shown in figure 1 there are many styles of fake news types and specially with low resourced languages it would be hard to define and detect the news within the huge amount of data that produced every single minute through deferent channels starting with social network sites (Al-Rabeeah et al., 2019), non-licensed websites, TV, and streaming (Allen et al., 2020).

Social media's influence and penetration have significantly altered the reach of false information. Its reach has increased thanks to the development of intelligent devices and extremely affordable internet (Al-Rabeeah et al., 2017). Even the most inaccessible areas in different places have access to smart phones and the internet in low resourced language areas. Even though these facilities have many advantages, they originated at a cost in the form of the immediate spread of bogus information alongside verifiable information (Shu et al., 2017b). The number of individuals use up social media and blogging has significantly amplified during the past twelve years (Ahmadi, S., Hassani., 2020). The amount of data that posts online are increasing day by day in the amount on comments, likes, blogs would increase the popularity of each published news if it is fake or real. (Khanam, Alwasel, Sirafi, & Rashid, 2021). fosters advancement in the fields of techniques and approaches in the veracity of these posts. There have been studies where phony news items have been automatically detected using machine learning. Additionally, only a few studies have used deep learning for automatic feature extraction in fake news detectors (Rodriguez et al., 2012).

Tokenization is the process of identifying text segment boundaries in natural language processing (NLP). More specifically, word tokenization and sentence tokenization refer to obtaining the boundaries of words and phrases, respectively. So, a tokenization system, commonly referred to as a lexical analyzer or tokenizer, divides a string of letters into tokens, of words or sentences (Kaplan, 2005).

2.RELATED WORK

This section presents a thorough analysis to comprehend the success ratio in finding fake news on lesser-resourced languages such as Arabic and Urdu. This is because they are like the Kurdish language in terms of their orientation, which is from right to left.

Authors of (Al-Yahya et al., 2021) show that transformer-based models perform better than neural network-based solutions, raising the F1 notch from 0.83 (best neural network-based model, GRU) to 0.95 (best transformer-based model, QARiB), and increasing accuracy by 16% in contrast to the finest neural network-based solutions. Then, they list the major research flaws in Arabic FND and suggest potential research instructions.

Furthermore, (Himdi et al., 2022) They develop a controlled machine learning prototype that classifies Arabic news articles centered on the authenticity of their framework to address the difficulties associated with news authentication. Additionally, they present the initial dataset of crowdsourced Arabic fake news pieces. They then develop a distinctive method of developing Arabic lexical wordlists and establish an Arabic Natural Language Processing program to accomplish textual characteristics extraction from the articles. The results of this investigation indicate enormous potential and superior performance compared to those of humans in the identical task (Veisi, H.,2020).

An advanced novel work was pursued by (Azad et al., 2021) where several classifiers are applied to the quantity once utilizing TF-IDF as a feature of variety. It includes two collections of news: the first set comprises crawling false news, while the second set contains text that has been altered from actual news. The results of the suggested research demonstrated that Support Vector Machine (SVM) outperformed the other methods on set 1 and achieved the maximum accuracy of 88.71 percent among the classifiers.

Hence, in the Urdu language (Humayoun, 2022) One of the models that they used in their study to check the news if it is fake or real obtained a high score in F1 factor in measuring the performance of their algorithm Macro score of 0.6674, which was above average than the competition's second-highest score. Lemmatization, Support Vector Machines (polynomial kernel degree 1), and the selection of the top 20K features from a total of 1.557 million features (which were generated by Word n-grams with n=1,2,3,4 and Char n-grams with n=2,3,4,5,6) were used to reach the result, The overall results show significant achievement in the study for low resourced languages.

To illustrate the research gap that we are targeting in our study actually to try to build a system they can detect any type of news in low resourced languages, in this study we will focus on Kurdish language, hence there is limited researches has been conducted on Kurdish language due to the limited resource available online to use in this matter.

3.METHODOLOGY

3.1 In this section, the proposed system architecture and its components are outlined and briefly explained. The system architecture is shown in the subsequent section. 3.1 System Architecture.

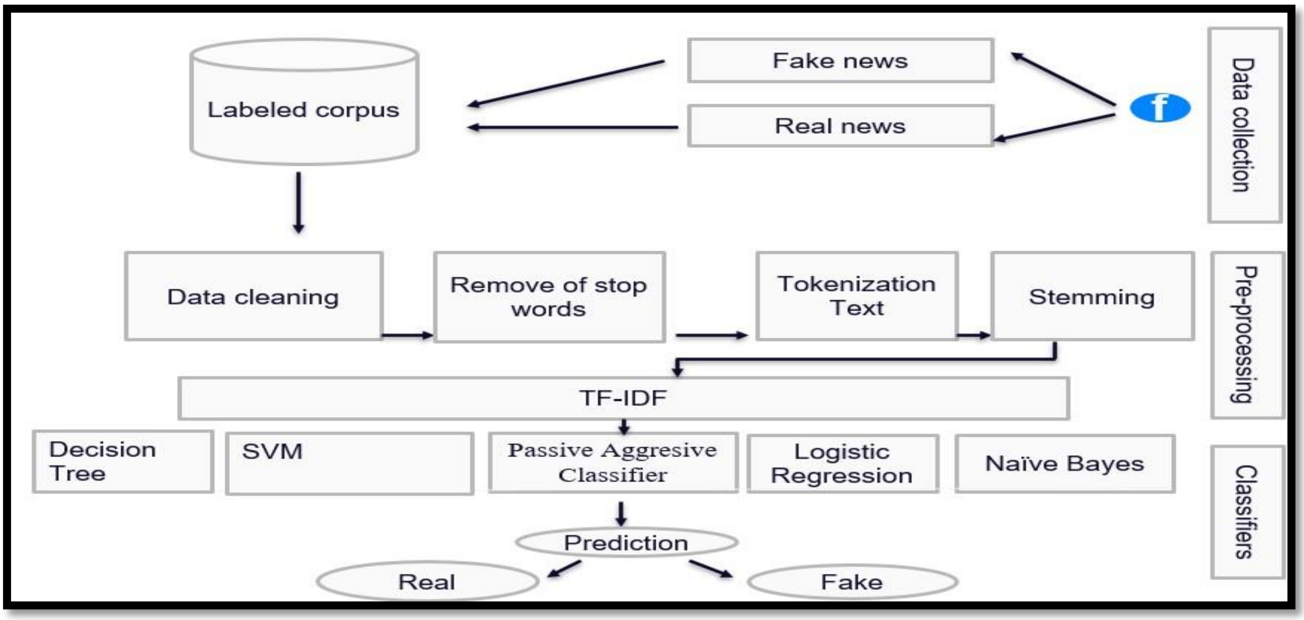


Figure 2: System Architecture

3.2 System Components

In this section the components of the system architecture are explained.

3.2.1 Dataset

As shown in the figure 3 the data were collected from different sources that authenticated for the real news and the fake news from Facebook social media platform.

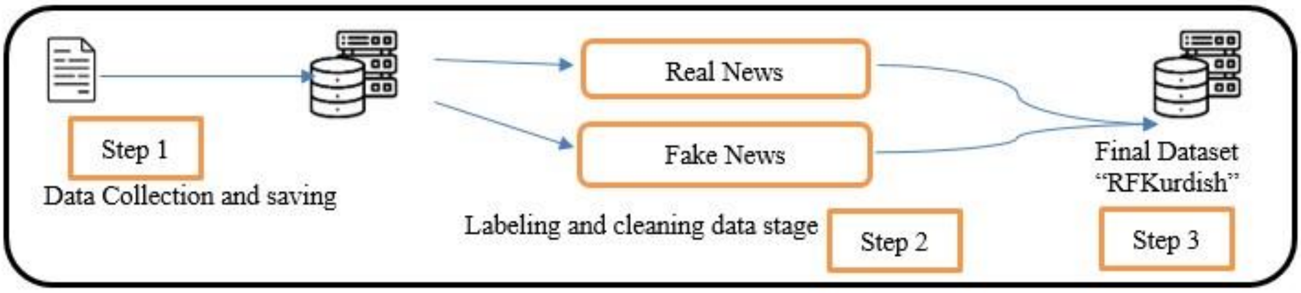
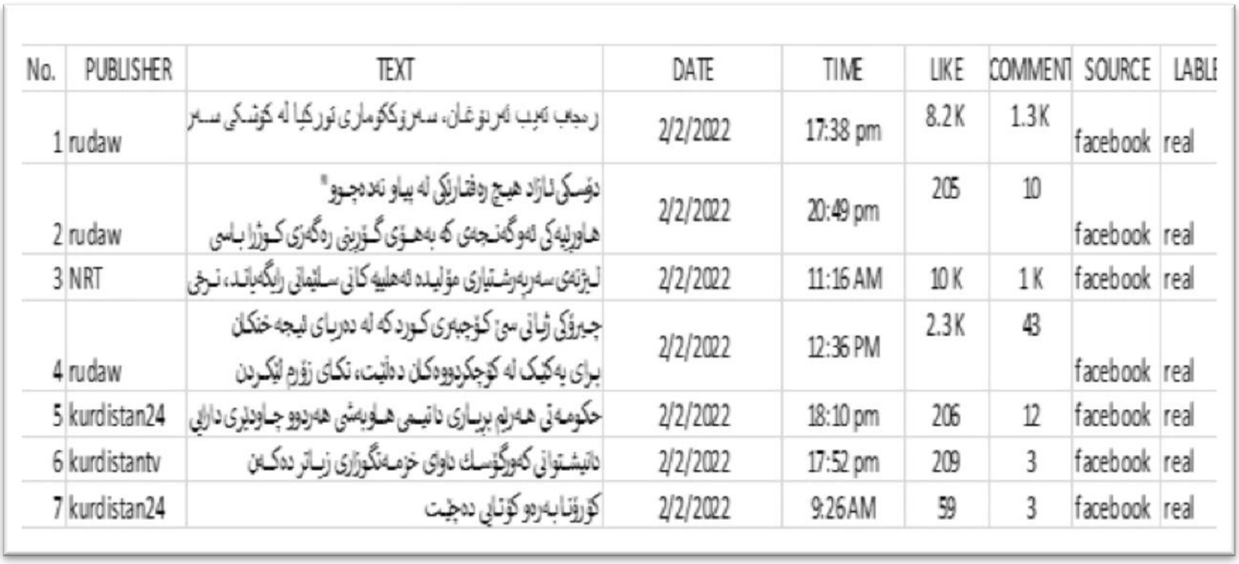


Figure 3: Data Collection process

So, the process of collecting data is consisting of the following steps:

1. Genuine Data collection: Manually, 6000 pieces of genuine news were collected from six reputable and authorized Facebook sites in Kurdistan: Rudaw, K24, NRT, KNN,Kurdistantv and Kurdsat news. These pages covered topics such as finance, economy, health, politics, culture, sports, and technology such as this is an example of our real news dataset.

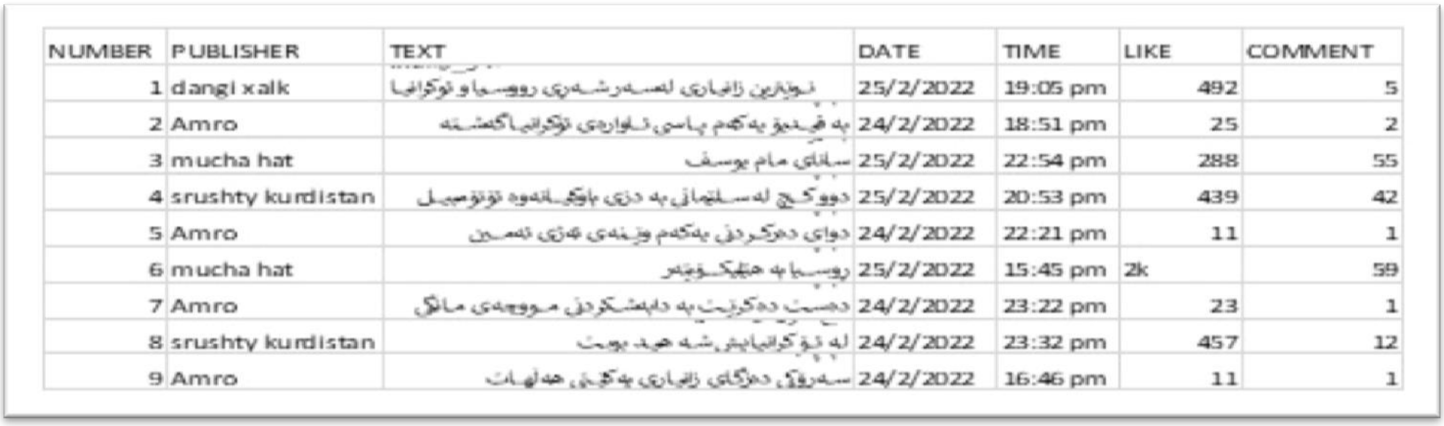


No.	PUBLISHER	TEXT	DATE	TIME	LIKE	COMMENT	SOURCE	LABLE
1	rudaw	رەجەب تەیب ئەردۆغان، سەرۆککۆماری تورکیا لە کۆشکی سەر	2/2/2022	17:38 pm	8.2K	1.3K	facebook	real
2	rudaw	دۆسکی ئازاد هیچ رەفتارێکی لە پیاو نەدەچوو" هاوڕێیەکی ئەوگەنجەی کە بەهۆی گۆڕینی رەگەزی کوژرا باسی	2/2/2022	20:49 pm	205	10	facebook	real
3	NRT	لیژنەی سەرپەرشتیاری مۆلیدە ئەهلییەکانی سلێمانی ڕایگەیاند، نرخی	2/2/2022	11:16 AM	10 K	1 K	facebook	real
4	rudaw	چیرۆکی ژیانی سێ کۆچبەری کورد کە لە دەریای ئینجە خنکان برای یەکێک لە کۆچکردووەکان دەڵێت، تکای زۆرم لێکردن	2/2/2022	12:36 PM	2.3K	43	facebook	real
5	kurdistan24	حکومەتی هەرێم بڕیاری دانیمی هەڵوێستی هەردوو چاودێری دارایی	2/2/2022	18:10 pm	206	12	facebook	real
6	kurdistantv	دانیشتوانی کەرگوسک داوای خزمەتگوزاری زیاتر دەکەن	2/2/2022	17:52 pm	209	3	facebook	real
7	kurdistan24	کۆرۆنا بەرەو کۆتایی دەچێت	2/2/2022	9:26 AM	59	3	facebook	real

Figure 4: Real News sample

2.Fake data collection: Manually, 6000 pieces of fake news were collected from non-legitimate Facebook pages that match the below conditions to evaluate fake pages

- The post's title isn't suitable for its subject concern.
- It features an unreliable link.
- An idealistic picture, and questionable sources. Such as this is an example of our fake news dataset.



NUMBER	PUBLISHER	TEXT	DATE	TIME	LIKE	COMMENT
1	dangi xalk	نوێترین زانیاری لەسەرشەڕی ڕووسیا و ئۆکرانیا	25/2/2022	19:05 pm	492	5
2	Amro	بە ڤیدیۆ یەکەم پاسی ئاوارەی ئۆکرانیا گەشتە	24/2/2022	18:51 pm	25	2
3	mucha hat	سالای مام یوسف	25/2/2022	22:54 pm	288	55
4	srushty kurdistan	دوو کچ لەسلێمانی بە دزی باوکیانەوە ئۆتۆمبیل	25/2/2022	20:53 pm	439	42
5	Amro	دوای دەرکردنی یەکەم وێنەی هەژی ئەمین	24/2/2022	22:21 pm	11	1
6	mucha hat	ڕووسیا بە هێلیکۆپتەر	25/2/2022	15:45 pm	2k	59
7	Amro	دەست دەکرێت بە دابەشکردنی مووچەی مانگ	24/2/2022	23:22 pm	23	1
8	srushty kurdistan	لە ئۆکرانیایش شەهید بووت	24/2/2022	23:32 pm	457	12
9	Amro	سەرۆکی دەزگای زانیاری یەکێتی هەڵهات	24/2/2022	16:46 pm	11	1

Figure 5: Fake News sample

3. After all the data were collected and labeled the fake news and real news were combined resulting the final dataset that we call it “RFkurdish” (shekhan eman 2022) that consists of 12000 sample of news mixed as 6000 reals plus 6000 fake.

3.2.2 Data pre-processing

In the pre-processing stage we will do the following

Prior to feeding text data to classifiers, processing is essential in NLP to enhance the superiority of the text data by removing extraneous information. Kurdish Language Processing Toolkit KLPT (Ahmadi, 2020) was employed in this work to achieve the following goals:

Data polishing, UTF-8 encoding, and standardization

It is crucial to remove special characters like, @, percent, &..., URLs, words in other languages, emojis, and unnecessary spaces from text data in order to improve its excellence and assure the accuracy of statistical analysis in order to have a relevant analysis. then change the text's encoding to UTF-8.

Tokenization

The goal is to break up the news into a succession of single words divided by white space because textual materials in real language are typically constructed of long, convoluted, and deformed sentences.

Removal of Stop Words

Use a list of Kurdish stop words that contains 240 stop words to get rid of unnecessary words like connectors, articles, and pronouns (Mustafa & Rashid, 2018).

Stemming

Process sending a word to its root form or source to improve the implementation of text extraction.

3.2.3 Feature Extraction

Feature extraction is a significant stage to select the appropriate feature sets. Term Frequency-Inverse Document Frequency **(TF-IDF)** has been used commonly in literature to transform the text into numerical values which can be fed to a machine learning model for processing. It provides insights about the less relevant and more relevant words in a document. It is considered a simple technique.

3.2.4 Classifiers

This stage of the system architecture, five machine learning methods are applied for predictive modelling. These five classifiers are Support Vector Machine (SVM), Decision Trees (DT), Naive Bayes (NB), Passive-Aggressive Classifier (PAC), and Logistic Regression (LR). In this study those classifiers were proposed in order to eliminate classifier bias in predictive modelling by picking machine learning methods. So as a result, after we apply all the methods, we will choose the three highest accuracy classifiers to represent is it real or fake data in the same time showing different factors of each method to specify the efficiency of the work as shown in the result and discussion.

3.2.5 Measurement factors

In this section we will discuss the factors that used to measure the performance of the classifiers in the proposed system.

A. Accuracy: One parameter for assessing classification models is accuracy. The percentage of predictions that our model correctly predicted is known as accuracy as shown in the equation.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Where TP is true positive, and TN is true negative, and FP is false positive and lastly FN is false negative

B. Precision: is computed as the sum of true positives across all classes divided by the sum of true positives and false positives across all classes in an imbalanced classification issue with more than two classes.

$$Precision = \frac{TP}{TP + FP}$$

C. Recall: is measured as the sum of true positives across all classes divided by the sum of true positives and false negatives across all classes in an imbalanced classification issue with more than two classes.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

D. F1-score: By calculating the harmonic mean of a classifier's precision and recall, the F1-score integrates both into a single metric. It mainly used to compare the effectiveness of two classifiers. Assume classifiers A and B have higher recall and precision, respectively.

$$F1 = 2 * (\frac{precision * recall}{precision + recall})$$

C. Support is representing the ability of using each algorithm with the datasets that we have.

Rule X ➡Y

$$Support = \frac{\mathrm{Frequency(X, Y)}}{\mathrm{N}}$$

$$Confidence = \frac{Frequency(X,Y)}{\mathrm{Frequency(X)}}$$

$$Lift = \frac{\mathrm{Support}}{\mathrm{Support(X)} * \mathrm{Support(Y)}}$$

Where X repented the true values and Y the false values and TP represented as True positive and TF represented as true false and FP as false positive and FN as false negative.

4.RESULTS AND DISCUSSIONS

By using python Spyder to run our proposed model the outcomes of our experiment on utilizing the five classifiers on our dataset utilizing the accuracy, precision, recall, support, and F1 score metrics to assess the classifiers and confusion matrix to show the percentage of each classifier as shown in figure 6.

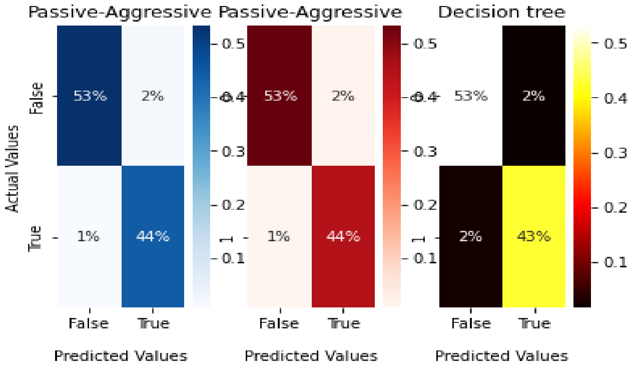
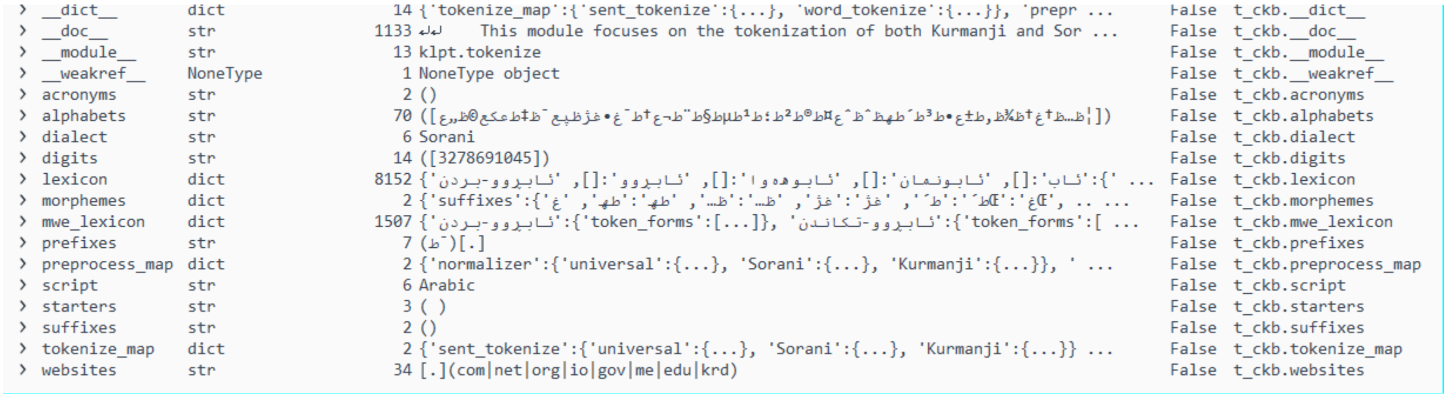


Figure 6: Confusion Matrix results

And the memory used for running that process for check a sentence were 86% and execution time equal to 3 to 5 seconds with the following software version used (python 3.9.7) and hardware intel core i7 8th generation and 8 GB of RAM.

In the next figure 7we are showing a sample of how the proposed system will handle the data step by step and train it according to the percentage that we will specify in the code, the text the used for the test were as (هێرشێکی گەورەی مووشەکی بۆ سەر کۆگا نەوتییەکانی شاری ئێۆدیسا لە ئۆکراینا)

So, the trained data will be run first then will check the text if it is real or fake data.



```
›  __dict__         dict           14 {'tokenize_map':{'sent_tokenize':{...}, 'word_tokenize':{...}}, 'prepr ...     False  t_ckb.__dict__
›  __doc__          str          1133 لەلە     This module focuses on the tokenization of both Kurmanji and Sor ...     False  t_ckb.__doc__
›  __module__       str            13 klpt.tokenize                                                                   False  t_ckb.__module__
›  __weakref__      NoneType        1 NoneType object                                                                 False  t_ckb.__weakref__
›  acronyms         str             2 ()                                                                              False  t_ckb.acronyms
›  alphabets        str            70 ([ع„ظ©عکعظ‡ظ¯غژظپع•غ¯ظ†ع¬ظ¨ظ§ظµظ¹ظ؛ظ²ظ®ظ¤ع^ظ^ظهظ´ظ³ظ•ع±ظ,ظ¾ظ†غ†ظ…ظ¦])                                 False  t_ckb.alphabets
›  dialect          str             6 Sorani                                                                          False  t_ckb.dialect
›  digits           str            14 ([3278691045])                                                                  False  t_ckb.digits
›  lexicon          dict         8152 {'ئابڕوو-بردن' ,[]:'ئابڕوو' ,[]:'ئابوهەوا' ,[]:'ئابونمان' ,[]:'ئاب':{' ...     False  t_ckb.lexicon
›  morphemes        dict            2 {'suffixes':{'غ' ,'ظه':'ظه' ,'ظ…':'ظ…' ,'غژ':'غژ' ,'ظ´':'ظ´' ,'ظŒ':'غŒ', .. ...     False  t_ckb.morphemes
›  mwe_lexicon      dict         1507 {'ئابڕوو-بردن':{'token_forms':[...]}, 'ئابڕوو-تکاندن':{'token_forms':[ ...    False  t_ckb.mwe_lexicon
›  prefixes         str             7 (ظ¯)[.]                                                                          False  t_ckb.prefixes
›  preprocess_map   dict            2 {'normalizer':{'universal':{...}, 'Sorani':{...}, 'Kurmanji':{...}}, ' ...     False  t_ckb.preprocess_map
›  script           str             6 Arabic                                                                          False  t_ckb.script
›  starters         str             3 ( )                                                                             False  t_ckb.starters
›  suffixes         str             2 ()                                                                              False  t_ckb.suffixes
›  tokenize_map     dict            2 {'sent_tokenize':{'universal':{...}, 'Sorani':{...}, 'Kurmanji':{...}} ...     False  t_ckb.tokenize_map
›  websites         str            34 [.](com|net|org|io|gov|me|edu|krd)                                              False  t_ckb.websites
```

Figure 7: Tokenized process

Table 1. shows the performance evaluation of five classifiers in KLPT model on our dataset that includes fake news and real news.

Table 1: Results of RFkurdish dataset when its fake

Classifier	Accuracy	Precision	recall	F1-score	Support
Logistic Regression:	0.91	0.99	0.83	0.90	335
Naive-Bayes:	0.88	0.83	0.95	0.88	335
Passive-Aggressive Classifier:	0.93	0.95	0.91	0.93	335
Decision Tree:	0.90	0.91	0.88	0.90	335
SVM Classifier:	0.91	1.00	0.83	0.91	335

The table shows that the passive aggressive have the highest accuracy among the five classifiers with accuracy of 93% while the support vector machine and logistic regression both came in the second highest as 91% and fourth Decision tree with 90% while at the end Naïve-Bayes with 88% and when we check the second column we will see the highest precision is SVM but it's less than the PAC in the other factors so cannot be considered as the top classifier in this case therefore the PAC as in average have the highest score in all the factors.

Table 2 will illustrate the results when it's True as 1 or real data.

Table 2: Results of RFkurdish dataset when its real data

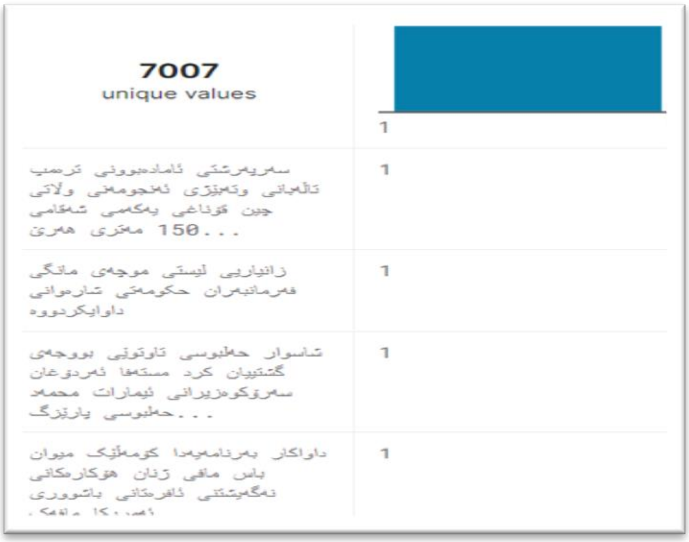
Classifier	Accuracy	Precision	recall	F1-score	support
Logistic Regression:	0.91	0.85	0.99	0.92	332
Naive-Bayes:	0.88	0.94	0.80	0.86	332
Passive-Aggressive Classifier:	0.93	0.91	0.95	0.93	332
Decision Tree:	0.90	0.89	0.92	0.90	332
SVM Classifier:	0.91	0.86	1.00	0.92	332

The PAC classifier also dominated with the highest results almost in all factors rather than the accuracy, as shown in the second table. If we do comparison of our results with those of previous studies that have been done to detect fake news by using different datasets that were collected by them as the real data and for the fake were generated using python code, So, to make sure of our work performance we have used their dataset as they have shared it online and we have got the following results as shown in table 3.

Table 3: (Azad et al., 2021) comparing the results

Classifier	Accuracy	Accuracy of our work	Precision	Precision of our work	Recall	Recall of our work	F1-score	F1- score of our work
Logistic Regression:	0.88	0.91	0.88	0.85	1.00	0.99	0.93	0.92
Naive-Bayes:	0.88	0.88	0.88	0.94	1.00	0.80	0.93	0.86
Passive-Aggressive Classifier	0.75	0.93	0.86	0.91	0.86	0.95	0.86	0.93
Decision Tree:	0.50	0.90	0.80	0.89	0.57	0.92	0.67	0.90
SVM Classifier	0.88	0.91	0.88	0.86	1.00	1.00	0.93	0.92

By testing the dataset on our code that they build it using a tool to retrieve news from social media sites and here is an example of their dataset in the figure 6 that they published online in Kaggle database,



7007 unique values	1
سەرپەرشتی ئامادەبوونی ترمپ تاڵەبانی وتەبێژی ئەنجومەنی وڵاتی چین قۆناغی یەکەمی شەقامی 150 مەتری هەری...	1
زانیاریی لیستی موچەی مانگی فەرمانبەران حکومەتی شارەوانی داوایکردووە	1
شاسوار حەلبوسی تاوتوێی بووجەی گشتییان کرد مستەفا ئەردوغان سەرۆکوەزیرانی ئیمارات محمەد حەلبوسی پارێزگ...	1
داواکار بەرنامەیەدا کۆمەڵێک میوان باس مافی ژنان هۆکارەکانی نەگەیشتنی ئافرەتانی باشووری	1

Figure 6: sample of Azad dataset

To highlight the improvement in our model than previous studies we achieved more accuracy and faster results with bigger amount of news as we mentioned earlier as 12000 fake and real news divided by half equally.

For their proposed model they have test 30% of the data and 70% for training and they got the results accordingly as shown in the above table, we got the same result they got for the SVM and lower results for the other classifiers and that would be justified for the language different because there is different dialogs in the Kurdish language so this could be from the major challenges could face the researchers during the studies and the hardest thing to do is collecting the data from the internet as we mentioned earlier in this paper.

5.CONCLUSIONS

We created a framework utilizing the KLPT model and applied five different classifiers in this paper to discuss the use of various classifiers for detecting fake news in Kurdish. The results showed that the PAC classifier outperformed the other classifiers with 93% accuracy for the "RFkurdish" dataset, with the SVM and LR coming in second with 91%. Regarding identifying fake news in a language like Kurdish that has few resources, there are still a lot of unresolved issues and research gaps. In the future, we want to focus on using ensemble methods and sentiment analysis to spot fake news in the Kurdish language by building an online system that could be connected to live dataset that will be updated automatically be getting the news directly online from different platforms and in the same time compared it to check the news in very fast response system online on the spot.

References:

Ahmadi, S., 2020, November. KLPT–Kurdish Language Processing Toolkit. In *Proceedings of Second Workshop for NLP Open-Source Software (NLP-OSS)* (pp. 72-84).

Ahmadi, S., Hassani, H., & Abedi, K. (2020, May). A corpus of the Sorani Kurdish folkloric lyrics. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)* (pp. 330-335).

Allen, J., Howland, B., Mobius, M., Rothschild, D. and Watts, D.J., 2020. Evaluating the fake news problem at the scale of the information ecosystem. *Science advances*, *6*(14), p. eaay3539

Al-Rabeeah, A.A.N. and Hashim, M.M., 2019. Social Network Privacy Models. *Cihan University-Erbil Scientific Journal*, *3*(2), pp.92-101.

Al-Rabeeah, A.A.N. and Saeed, F., 2017, May. Data privacy model for social media platforms. In *2017 6th ICT International Student Project Conference (ICT-ISPC)* (pp. 1-5). IEEE.

Al-Yahya, M., Al-Khalifa, H., Al-Baity, H., AlSaeed, D. and Essam, A., 2021. Arabic fake news detection:

comparative study of neural networks and transformer-based approaches. *Complexity*, *2021*.

Azad, R., Mohammed, B., Mahmud, R., Zrar, L. and Sdiqa, S., 2021. Fake News Detection in low-resourced languages "Kurdish language" using Machine learning algorithms. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(6), pp.4219-4225.

Bryar A. Hassan, Tarik A. Rashid, Seyedali Mirjalili, Performance evaluation results of evolutionary clustering algorithm star for clustering heterogeneous datasets, Data in Brief, Volume 36, 2021, 107044, ISSN 2352-3409,https://doi.org/10.1016/j.dib.2021.107044.

Himdi, H., Weir, G., Assiri, F. and Al-Barhamtoshy, H., 2022. Arabic fake news detection based on textual analysis. *Arabian Journal for Science and Engineering*, pp.1-17.

Humayoun, M., 2022. The 2021 Urdu Fake News Detection Task using Supervised Machine Learning and Feature Combinations. *arXiv preprint arXiv:2204.03064*.

Khalifa, M. and Hussein, N., 2019, January. Ensemble learning for irony detection in Arabic tweets.

Khanam, Z., Alwasel, B. N., Sirafi, H., & Rashid, M. (2021). Fake News Detection Using Machine Learning Approaches. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 012040. https://doi.org/10.1088/1757-899x/1099/1/012040

Martin Forst and Ronald M Kaplan. 2006. The importance of precise tokenizing for deep grammars. In LREC, pages 369–372.

Mustafa, A. M., & Rashid, T. A. (2018). Kurdish stemmer pre-processing steps for improving information retrieval. *Journal of Information Science*, *44*(1), 15–27. https://doi.org/10.1177/0165551516683617

Oshikawa, R., Qian, J. and Wang, W.Y., 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

Rodriguez, M., Peterson, R.M. and Krishnan, V., 2012. Social media's influence on business-to-business sales performance. *Journal of Personal Selling & Sales Management*, *32*(3), pp.365-378.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, *19*(1), 22-36.

Tony McEnery and Andrew Wilson. 2003. Corpus linguistics. The Oxford handbook of computational linguistics, pages 448–463.

Veisi, H., MohammadAmini, M., & Hosseini, H. (2020). Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus. *Digital Scholarship in the Humanities*, *35*(1), 176-193.