# Fake News Detection in Arabic Tweets during the COVID-19 Pandemic

Ahmed Redha Mahlous[1]
College of Computer and Information Sciences
Prince Sultan University
Riyadh, Saudi Arabia

Ali Al-Laith[2]
Center for Language Engineering - KICS
University of Engineering and Technology
Lahore, Pakistan

***Abstract*—In March 2020, the World Health Organization declared the COVID-19 outbreak to be a pandemic. Soon afterwards, people began sharing millions of posts on social media without considering their reliability and truthfulness. While there has been extensive research on COVID-19 in the English language, there is a lack of research on the subject in Arabic. In this paper, we address the problem of detecting fake news surrounding COVID-19 in Arabic tweets. We collected more than seven million Arabic tweets related to the corona virus pandemic from January 2020 to August 2020 using the trending hashtags during the time of pandemic. We relied on two fact-checkers: the France-Press Agency and the Saudi Anti-Rumors Authority to extract a list of keywords related to the misinformation and fake news topics. A small corpus was extracted from the collected tweets and manually annotated into fake or genuine classes. We used a set of features extracted from tweet contents to train a set of machine learning classifiers. The manually annotated corpus was used as a baseline to build a system for automatically detecting fake news from Arabic text. Classification of the manually annotated dataset achieved an F1-score of 87.8% using Logistic Regression (LR) as a classifier with the n-gram-level Term Frequency-Inverse Document Frequency (TF-IDF) as a feature, and a 93.3% F1-score on the automatically annotated dataset using the same classifier with count vector feature. The introduced system and datasets could help governments, decision-makers, and the public judge the credibility of information published on social media during the COVID-19 pandemic.**

***Keywords*—Fake news; Twitter; social media; Arabic corpus**

## I. Introduction

The rise of social networks such as Facebook, Twitter, and many others has enabled the rapid spread of information. Any user on social media can publish whatever they want without considering the truthfulness and reliability of the published information, which introduces challenges in information reliability assurance. Twitter is one of the most popular social media platforms. It is designed to allow users to send information as short texts, known as tweets, with no more than 280 characters, and each user on Twitter can follow as many accounts as he or she wants. Nowadays, and with the outbreak of the COVID-19 pandemic, millions of tweets are generated daily, which has caused some adverse effects that impact individuals and society. For example, the spread of misinformation about COVID-19 symptoms may harm people [1]. For instance, it could be anxiety-inducing for a person who experiences COVID-19 like symptoms even if they have not been infected with the virus. The terms fake news and misinformation are closely related and are often used interchangeably. Authors in [2] defined rumors as: "a hypothesis offered in the absence of verifiable information regarding uncertain circumstances that are important to those individuals who are subsequently anxious about their lack of control resulting from this uncertainty." Another definition presented in [3] is: "unverified and instrumentally relevant information statements in circulation that arise in contexts of ambiguity, danger or potential threat, and that function to help people make sense and manage risk."

Detecting fake news in English tweets is an active research area and many studies and datasets have been published during the COVID-19 pandemic [4]. In Arabic, fake news detection is fairly new and has a long way to go to reach the level achieved in other languages, especially English. Therefore, the fight against fake news requires a system that automatically assists in verifying the truthfulness of shared information about the COVID-19 pandemic on social media. Fake news detection is a very challenging task, especially with the lack of available datasets related to the pandemic. An automated fake news detection system is necessary by utilizing human annotation, machine/deep learning, and Natural Language Processing techniques [5]. These techniques help to determine whether a given text is fake news or not by comparing the text with some pre-known corpora that contain both fake and truthful information [6].

In this paper, we address the problem of fake news detection on Twitter during the COVID-19 pandemic period. Our focus is to build a manually annotated dataset for fake news detection from Twitter's social media platform. We rely on fact-checking sources to manually annotate a sample dataset. The consideration of these fact-checking sources could help in reducing the spread of misinformation [7], [8], [9], [10]. As manual annotation is expensive and time-consuming [11], we also developed a system to expand the manually annotated dataset by automatically annotating a large and unlabeled dataset. We use a supervised learning classification to train and test both the manually and automatically annotated datasets to ensure the quality of our annotation. We use six different machine learning algorithms, four different features with each algorithm, and three pre-processing techniques. The rest of the paper is organized as follows. In Section 2, we cover related work. Section 3 presents our methodology to annotate and automatically detect fake news related to COVID-19. In Section 4, we present the results and discussion. Finally, the conclusion and future work are presented in Section 5.

## II. Related Work

Recently, many works have been done to tackle the issue of detecting fake news, rumors, misinformation, or disinformation in social media networks. Most of these studies can be categorized into supervised and unsupervised learning approaches. Moreover, there are fewer works that tackled the problem using semi-supervised techniques.

For the supervised approach, a system based on machine learning techniques for detecting fake news or rumors in the Arabic language from social media during the COVID-19 pandemic is presented in [12]. The authors collected one million Arabic tweets using Twitter's Streaming API. The collected tweets were analyzed by identifying the topics discussed during the pandemic, detecting rumors, and predicting the source of the tweets. A sample of 2,000 tweets was labeled manually into false information, correct information, and unrelated. Different machine learning classifiers were applied, including Support Vector Machine, Logistic Regression, and Naïve Bayes. They obtained 84% accuracy in identifying rumors. The limitations of this research include the unavailability of the dataset, and the fact that it relies on a single source of rumors: the Saudi Arabian Ministry of Health.

Identifying breaking news rumors on Twitter has been proposed in [13]. The authors built a word2vec model and an LSTM-RNN model to detect rumors from news published on social media. The proposed model is capable of detecting rumors based on a tweet's text, and experiments showed that the proposed model outperforms state-of-the-art classifiers. As rumors can be deemed later to be true or false, their model is unable to memorize the facts across time; it only looks at the tweet at the current time. Detecting rumors from Arabic tweets using features extracted from the user and content has been proposed in [14]. The authors obtained rumors and non-rumors topics from anti-rumors and Ar-Riyadh websites. More than 270K tweets were collected, containing 89 and 88 rumour and non-rumour events, respectively. A supervised Gaussian Naïve Bayes classification algorithm reported an F1-score of 78.6%. This research's limitation is that the proposed dataset is not verified using any of the benchmark datasets.

In [15], a supervised learning approach for Twitter credibility detection is proposed. A set of features including content-based and source-based features, were used to train five machine learning classifiers. The Random Forests classifier outperformed the other classifiers when used with a combined set of features. A total of 3,830 English tweets were manually annotated with credible or non-credible classes. The textual features were not studied to examine their impact on credibility detection. Another supervised machine learning approach was proposed in [16] to detect rumors from business reviews. A publicly available dataset was used to conduct rumour detection experiments. Different supervised learning classifiers were used to classify business reviews. The experimental results showed that the Naïve Bayes classifier achieved the highest accuracy and outperformed three classifiers, namely, the Support Vector Classifier, K-Nearest Neighbors, and Logistic Regression. This work's limitation is the small size of the dataset used to train machine learning classifiers.

Detection of fake news using n-gram analysis and machine learning techniques was proposed in [17]. Two different feature extraction techniques and six machine learning algorithms were investigated and compared based on a dataset from political articles that were collected from Reuters.com and kaggle.com for real and fake news. Another Arabic corpus for the task of detecting fake news on YouTube is presented in [18]. The authors introduced a corpus that covered topics most concerned by rumors. More than 4,000 comments were collected to build the corpus. Three different machine learning classifiers (Support Vector Machine, Decision Tree, and Multinomial Naïve Bayes) were used to differentiate between rumour and non-rumour comments with the n-gram TF-IDF feature. The SVM classifier achieved the highest results. Authors in [19] proposed identifying fake news on social media. They used several pre-processing steps on the textual data, and then used 23 supervised classifiers with the TF weighting feature. The combined text pre-processing and supervised classifiers were tested on three different real-world English datasets, including BuzzFeed Political News, Random Political News, and ISOT Fake News.

An automatic approach to detecting fake news from Arabic and English tweets using machine learning classifiers has been proposed in [4]. The authors developed a large and continuous dataset for Arabic and English fake news during the COVID-19 pandemic. Information shared on official websites and Twitter accounts were considered a source of real information. Along with the data collected from official websites and Twitter accounts, they also relied on various fact-checking websites to build the dataset. A set of 13 machine learning classifiers and seven other feature extraction techniques were used to build fake news models. These models were used to automatically annotate the dataset into real and fake information. The dataset was collected for 36 days, from the 4th of February to the 10th of March 2020.

A large corpus for fighting the COVID-19 infodemiconsocial media has been proposed in [11]. The authors developed a schema that covers several categories including advice, cure, call for action, or asking a question. They considered such categories to be useful for journalists, policymakers, or even the community as a whole. The collected dataset contains tweets in Arabic and English. Three classifiers were used to perform classification experiments using three input representations: word-based, FastText, and BERT. The authors only made 210 of the classified tweets public.

Two Arabic corpora have also been constructed, without manual annotation. In [20], more than 700,000 Arabic tweets were collected from Twitter during the COVID-19 period. The corpus covers prevalent topics discussed in that period and is publicly available to enable research under different domains such as NLP, information retrieval, and computational social media. They used the Twitter API to collect the tweets on a daily basis, covering the period from January 27, 2020, to March 31, 2020.

The second corpus is presented in [21]. The tweets were collected during the period of the COVID-19 pandemic to study the pandemic from a social perspective. The corpus was developed to identify information influencers during the month of March 2020, and contains nearly four million tweets. Different algorithms were used to analyze the influence of information spreading and compare the ranking of users.

For fake news detection in other languages, there are many corpora that are publicly available to tackle the spread of false information. A multilingual cross-domain fact-checking news dataset for COVID-19 has been introduced in [22]. The collected dataset covered 40 languages and relies on fact-checked articles from 92 different fact-checking websites to manually annotate the dataset. The dataset is available on GitHub . Another publicly available dataset called "TweetsCOV19" was introduced in [23]. This dataset contains more than eight million English tweets about the COVID-19 pandemic. The dataset can be used for training and testing a wide range of NLP and machine learning methods and is available online. A novel Twitter dataset is presented in [24], which was developed to characterize COVID-19 misinformation communities. Authors categorized the tweets into 17 classes, including fake cure, fake treatment, and fake facts or prevention. They performed different tasks on the developed dataset, including identifying communities, network analysis, bot detection, sociolinguistic analysis, and vaccination stance. This study's limitations are that only one person performed annotation, the analyses are correlational and not causational, and the collected data covered a short period of only three weeks. MM-COVID is a multilingual and multidimensional fake news data repository [25]. The dataset contains 3981 fake and 7192 genuine news contents from English, Spanish, Portuguese, Hindi, French, and Italian. The authors explored the collected dataset from different perspectives including social engagements and user profiles on social media.

Sentiment analysis has also been used in fake news detection has also been facilitated. In [26], the authors used sentiment analysis to eliminate neutral tweets. They claimed that tweets related to fake news are more negative and have strong sentiment polarity in comparison with genuine news. The main issue in using this approach to detect fake news from Arabic text is the lack of Arabic sentiment resources, including sentiment lexicons and corpora [27]. Testing whether emotions play a role in the formation of beliefs in online political misinformation is presented in [28]. The authors explore emotional responses as an under-explored mechanism of belief in political misinformation. Understanding emotions helps in different domains including capturing the public's sentiments about social events such as the spreading of misinformation on social media [29].

Text classification using machine/deep learning provides a good results over many NLP applications including, sentiment analysis [30], [31], emotion detection [32], hate speech detection [33], sarcasm detection [34], and other applications.

To summarize, most of the existing datasets target the English language, with only a few targeting Arabic. Furthermore, most of the Arabic datasets related to COVID-19 are published without annotation. Datasets that are annotated were annotated automatically and collected during a short period of time. Additionally, not all of these datasets are publicly available. In this research, we address these issues by employing three annotators to manually perform the annotation task.

## III. Methodology

Fig. 1 presents the architecture of the proposed fake news detection system. In the first step of the framework, we collect data from Twitter using the Twitter Streaming API. In the second step, we perform the extraction of tweets which discuss rumors or fake news topics during the pandemic, annotate a small sample of tweets manually, and develop a system to annotate a large dataset of unlabeled tweets automatically.

In the last step, we store the dataset in a database and use it to accomplish our experiments and analysis. This research intends to build an Arabic fake news corpus that can be used for analyzing the spread of fake news on social media during the COVID-19 pandemic. To address this need, we perform the following four steps: 1) data collection, 2) rumor/misinformation keyword extraction, 3) data preprocessing, and 4) fake news annotation.

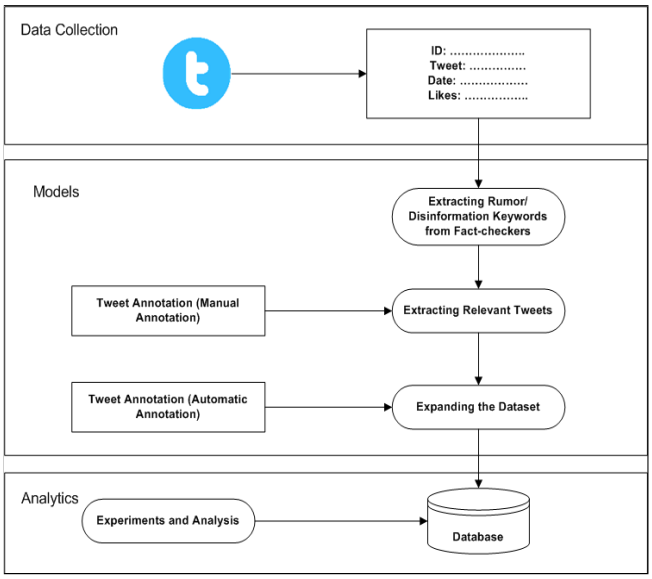


Fig. 1. Fake New Detection Architecture.

### A. Data Collection

In this section, we describe the process of data collection from Twitter. In the first instance, we prepared a list of hashtags that appeared during the COVID-19 outbreak, as shown in Table I. Armed with the Tweepy Python library and using Twitter's API, we proceeded to collect Arabic tweets related to COVID-19 from January 1, 2020, until May 31, 2020. We then searched for tweets containing one or more of the defined hashtags in the tweet's text. This step allowed us to collect more than seven million unique tweets. After applying some filters such as removing the short and repeated tweets, the remaining tweets are 5.5 million tweets. However, as some of the collected tweets were irrelevant, we decided to keep only those tweets relevant to the COVID-19 pandemic and containing fake news keywords.

### B. Fake News Keywords Extraction

To collect a list of keywords relevant to the rumours circulating during the pandemic, we used two sources:

- Agence France-Presse (AFP)[1] with its newly formed health investigation team which have the responsibility

[1]https://factuel.afp.com/ar/CORNA%20COMPILATION%202-20

TABLE I. LIST OF HASHTAGS USED TO COLLECT THE DATASET

| # | Hashtag | English Translation |
|---|---|---|
| 1 | #كورونا | Corona |
| 2 | #فيروس_كورونا | Corona virus |
| 3 | #كورونا_المستجد | New Corona |
| 4 | #كوفيد_١٩ | COVID 19 |
| 5 | #الفيروس_التاجي | Corona virus |
| 6 | #الحجر_المنزلي | Home Quarantine |
| 7 | #الحجر_الصحي | Quarantine |
| 8 | #التباعد_الاجتماعي | Social distancing |
| 9 | #كلنا_مسؤول | We are all responsible |
| 10 | #ابقوا_في_منازلكم | Stay in your homes |

of dealing with large amounts of fake news in various languages and indicating their error or inaccuracy.

- The Anti-Rumours Authority (No Rumours)[2], an independent project established in 2012 to address and contain rumours and sedition to prevent them from causing any harm to society.

After reading and analyzing rumours and misinformation circulated on social media using the above-mentioned sources, a list of 40 keywords was extracted and used to prepare our dataset, as shown in Table II. These keywords cover a variety of topics associated with fake news, rumours, racism, unproven cure methods, false information. For example, there was a rumour that herbal tea is used to treat COVID-19. Another topic circulated was that Cristiano Ronaldo offered to transform his hotels into hospitals and give free treatment to COVID-19 patients. One alleged that the corona virus targets only those who have yellow-skin and Asian people to reduce population density. Other topics include the conversion of non-Muslims to Islam.

We extracted a corpus of more than 37,000unique tweets related to rumors and misinformation topics during the COVID-19 pandemic. The tweets were written by 24,117 users with an average of 1.5 tweets per user. Statistical information details about the corpus are presented in Table III.

### C. Data Pre-processing

We performed several text pre-processing steps based on the procedure described in [35] in order to sanitize the collected tweets before annotation and classification. Our dataset, which is a mixture of modern standard Arabic and dialectical Arabic, requires further filtering such as removing duplicated letters, strange words, and non-Arabic words. The following is a complete list of the steps performed:

- Removing mentions, hyperlinks, and hashtags.
- Removing non-Arabic and strange words.
- Text normalization.

TABLE II. LIST OF VERIFIED RUMORS AND MISINFORMATION TOPICS

| # | Keword | English Translation |
|---|---|---|
| 1 | شبكات الجيل الخامس | 5G networks |
| 2 | العجز الجنسي | Impotence |
| 3 | اكتشاف علاج | Discover a cure |
| 4 | درجة حرارة ٢٦ | A temperature of 26 |
| 5 | حبس النفس | Holding your breath |
| 6 | الغرغرة | Gargle |
| 7 | الموز | Banana |
| 8 | بخار الماء | Water vapor |
| 9 | تشخيص حالتك اونلاين | Online diagnosis of your condition |
| 10 | الشاي العشبي | Herbal tea |
| 11 | يلقون باموالهم | Throw in their money |
| 12 | يرمون اموالهم | Throw in their money |
| 13 | مجففات الايدي | Hand dryers |
| 14 | الساونا | Sauna |
| 15 | الرئيس يبكي | The president is crying |
| 16 | المياة ١٥ دقيقة | Water 15 minutes |
| 17 | يبصقون | Spit |
| 18 | الزنداني كورونا | Zindani Corona |
| 19 | الاويغور | Uyghurs |
| 20 | الأعشاب | Herbs |
| 21 | يدخلون الإسلام | They enter Islam |
| 22 | الرئيس الصيني | Chinese President |
| 23 | البشرة السوداء | Black skin |
| 24 | يعتنقون | Embrace |
| 25 | توزيع مصاحف | Distributing the Qur'an |
| 26 | صدام حسين | Saddam Hussein |
| 27 | مستشفيات رونالدو | Ronaldo Hospitals |
| 28 | الاذان غرناطة | Call to prayer in Granada |
| 29 | الاذان لندن | Call to prayer in London |
| 30 | هروب صيني الى اليمن | Chinese escape to Yemen |
| 31 | اسلمو | Entered Islam |
| 32 | جميع الاديان | All religions |
| 33 | تاسوكو | Tasuco |
| 34 | أوباما | Obama |
| 35 | بيل غيتس | Bill Gates |
| 36 | رفع اذان المانيا | Call to prayer in Germany |
| 37 | يصلون | They pray |
| 38 | متروك للسماء | Up to the sky |
| 39 | نقص الأكسجين | Lack of oxygen |
| 40 | المغرب يصدر الكمامات | Morocco exports masks |

- Removing punctuations and Arabic diacritics.
- Removing repeated characters which add noise and influence the mining process.

Two libraries were used to perform further pre-processing on the corpus text, including stemming and rooting. The first library is NLTK, which was used to perform stemming on the corpus text using ISRIStemmer4[3]. The second library is

[2]https://twitter.com/No_Rumors

[3]https://www.kite.com/python/docs/nltk.ISRIStemmer

Tashaphyne[4], which was used to get the root of each word in the corpus.

## IV. Corpus Annotation

### A. Manual Annotation

A sample of 2,500 tweets was manually annotated into fake or genuine classes. We developed a small application to facilitate the annotation process, as shown in Fig. 2. We involved three annotators in the annotation of the sample dataset. Two of the annotators performed annotation while the third was tasked with evaluating their output and resolving conflicts. We requested the annotators to read and understand the list of guidelines and informed them to skip tweets in which there are mixture of fake news and genuine topics and only annotate tweets that have a clear and distinct fake news topic. The following are the guidelines:

- Tweets are generally considered fake if one fake news topic is discussed in the tweet.
- Tweets are considered to not be fake if one fake news topic is discussed in the tweet, and the topic is negated.
- Tweets that contain a mixture of both fake and genuine news are skipped.

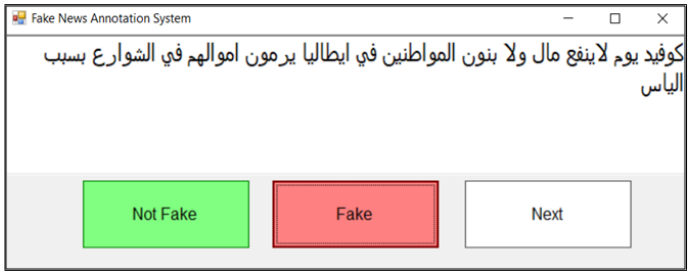


Fig. 2. Fake News Annotation Interface.

The annotation process resulted in a corpus containing 1,537 tweets (835 fake and 702 genuine), after excluding duplicated tweets, tweets that contain mixed fake and genuine news, and tweets where the fake news was meant as sarcasm. Statistical information about the manually annotated corpus is shown in Table III. We used Cohen's kappa coefficient to measure the inter-annotator agreement, obtaining a value of 0.91. Table IV shows an example of some annotated tweets.

### B. Automatic Annotation

Initially, we trained different machine learning classifiers on the manually annotated corpus and used the best performing classifier to automatically predict the fake news classes of remaining unlabeled tweets. The outcome of the prediction process is 34,529 tweets (19,582 fake and 19,582 genuine) as shown in Table III.

During the annotation process, the annotators found some tweets containing fake news keywords but carrying sarcasm. In this case, the annotators were requested to annotate them as genuine. Table V shows a sample of such tweets.

[4]https://pypi.org/project/Tashaphyne/

TABLE III. Corpus Statistics

Manually Annotated Corpus

| | Fake Tweets | Not Fake Tweets |
|---|---|---|
| Total Tweets | 835 | 702 |
| Total Words | 20,395 | 19,852 |
| Unique Words | 6,246 | 7,115 |
| Total Characters | 117,630 | 113,121 |

Automatically Annotated Corpus

| | Fake Tweets | Not Fake Tweets |
|---|---|---|
| Total Tweets | 19,582 | 14,947 |
| Total Words | 479,349 | 463,768 |
| Unique Words | 79,383 | 88,037 |
| Total Characters | 2,855,454 | 2,680,067 |

TABLE IV. Fake and Genuine Tweets Examples

| # | Tweet | Class |
|---|---|---|
| 1 | نفسي اكتب اشاعة علي فيروس كورونا اللي تخلي الناس مرزوعة في بيوتها اقول فيها كورونا يسبب الضعف الجنسي حتي لو شفيت منها.<br>I want to write a rumour about COVID-19 that would make people stay in their homes, and say that it causes impotence even if you recover from it. | Genuine |
| 2 | ربما مفبرك مثل كلام صدام حسين عن كورونا قبل اكثر من عشرين عام.<br>Perhaps fabricated like Saddam Hussein's video about COVID-19 more than 20 years ago. | Genuine |
| 3 | تحذير صيني فيروس كورونا يسبب الضعف الجنسي والعقم .<br>Chinese warning: COVID-19 causes impotence and sterility. | Fake |
| 4 | كورونا الفيروس مصنع لاصحاب البشرة الصفراء والاسيويين لتقليص الكثافة السكانية على الارض والدليل انه لم يصاب اي احد من ذوي البشرة السوداء.<br>COVID-19 is made for yellow-skinned people and Asians to reduce population density, and the evidence for that is that no black-skinned person has been infected. | Fake |

## V. Experiments

In this section, we present the results of the fake news classification after describing the employed feature extraction techniques, experimental setup, classifier model training, and evaluation measures.

### A. Feature Extraction

The next step after performing text pre-processing is to prepare the features to build classification models. To accomplish that, we used the following features:

- Count Vector: The text in our corpus was converted into a vector of term counts.
- Word-Level TF-IDF: Each term in our corpus is represented in a TF-IDF matrix.

TABLE V. Annotation confussion

| # | Tweet | English Translation |
|---|---|---|
| 1 | بالفيديو ابنة صدام حسين تحسم جدل فيديو والدها عن كورونا. | In the video, Saddam Hussein's daughter resolves the controversy of her father's video about COVID-19 |
| 2 | اقرا عن بيل غيتس ومسالة لقاح كورونا ومبلغ ال ٧٥ مليون | Read about Bill Gates, the issue of the COVID-19vaccine, and the 75 million[dollar] amount |

TABLE VI. STATISTICS IN EXPERIMENTAL DATASET

Manually Annotated Corpus

| | Fake Tweets | Not Fake Tweets | Total Tweets |
|---|---|---|---|
| Training | 668 | 562 | 1,230 |
| Testing | 167 | 140 | 307 |
| Total Tweets | 835 | 720 | 1,537 |

Automatically Annotated Corpus

| | Fake Tweets | Not Fake Tweets | Total Tweets |
|---|---|---|---|
| Training | 15,666 | 11,958 | 27,624 |
| Testing | 3,926 | 2,989 | 6,905 |
| Total Tweets | 19,582 | 14,947 | 34,529 |

- N-gram-Level TF-IDF: We used unigram, bigram, and trigram models in our experiments. We then represented these terms in a matrix containing TF-IDF scores.
- Character-Level TF-IDF: We represent TF-IDF character scores for each tweet in our corpus.

These features are used to train multiple classifiers in order to build machine learning models with the ability to decide the most probable category for new, unseen tweets.

### B. Experimental Setup

This section describes the experimental configurations used to perform the text classification task. We designed a set of experiments aiming to validate and ensure the quality of manually and automatically generated annotations. We also explored fake news detection as a binary classification problem (fake and genuine). The total tweets in our fake news dataset are 1,537 and 34,529 tweets in both manually and automatically annotated corpora, respectively. We divided both datasets into 80% for training and 20% for testing. Table VI shows detail about the manual and automatic annotated datasets.

Six machine learning classifiers were used to perform fake news classification for both datasets: Naïve Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest Bagging Model (RF), and eXtreme Gradient Boosting Model (XGB). The following are the hyper-parameters used with each classifier:

- NB: alpha=0.5
- LR: with default values
- SVM: c=1.0, kernel=linear, gamma=3
- MLP: activation function=ReLU, maximum iterations=30, learning rate=0.1
- RF: with default values
- XGB: with default values

### C. Models Training

Once the numerical form of the textual tweets was complete, the data frame containing the count vector, word-level TF-IDF, n-gram level TF-IDF, and character-level TF-IDF representations for each tweet in our corpus were used to train six different classifiers. We used scikit-learn, a Python library for classifier implementation and prediction of the classes of the unlabeled dataset. K-fold cross-validation was used to select the classifier that provides the highest results and shows the best ability to generalize. The collection was split into five-folds, four of which were used for training on each iteration, and the fifth for evaluation.

### D. Evaluation Measures

The evaluation was carried out using three measures: Precision, Recall, and F1-score as follows:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive)} \quad (1)$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative)} \quad (2)$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall)} \quad (3)$$

Where:

- True Positive: the number of fake tweets that are correctly predicted as fake tweets.
- True Negative: the number of genuine tweets that are correctly predicted as genuine tweets.
- False Positive: the actual class is genuine, but the predicted class is fake.
- False Negative: the actual class is fake, but the predicted class is genuine.

### E. Experimental Results

We present the experimental results on the Arabic fake news dataset. Six machine learning classifiers (NB, LR, SVM, MLP,RF, and XGB) were used to perform our experiments on the manually and automatically annotated datasets. We used count vector and TF-IDF vectorization (word-level, n-gram-level, and character-level) to train the classifiers. Precision, recall, and F1-score are the measures that have been used to evaluate each classifier using 5–fold cross-validation. Bold values indicate which setting yielded the best classification performance of fake tweets.

The results showed that using the LR classifier with the n-gram TF-IDF feature and without applying further pre-processing on the text (such as stemming or rooting) yielded a significantly better classification performance. The classifier gave an 87.8% F1-score classification result with the manually annotated corpus, as shown in Table VII. The same classifier, with the word count feature and without applying stemming or rooting, obtained the best classification performance when applied to the automatically annotated corpus, as shown in Table VIII. It achieved an F1-score of 93.3%.

As shown in Fig. 3, the highest precision value was obtained using the n-gram TF-IDF feature with the LR classifier (87.8%) and the count vector feature with the LR classifier (93.4%) on manually and automatically annotated corpora, respectively. The results obtained using raw text is better than with the corpus text after applying stemming and rooting. We can conclude that performing further pre-processing did not

TABLE VII. PRECISION (P), RECALL (R), AND F1-SCORE (F1) CLASSIFICATION RESULTS (MANUAL ANNOTATED DATASET)

| Feature | | Word Count | | | TF-IDF (word-level) | | | TF-IDF (n-gram-level) | | | TF-IDF (character-level) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Classifier | Measure | Raw Text | Stemming | Rooting | Raw Text | Stemming | Rooting | Raw Text | Stemming | Rooting | Raw Text | Stemming | Rooting |
| NB | P | 77.4 | 78.1 | 75.65 | 82.61 | 80.9 | 78.4 | 80.4 | 85.8 | 82.07 | 83.8 | 83.3 | 77.72 |
| | R | 77.1 | 77.5 | 74.68 | 73.38 | 75.6 | 70.13 | 74 | 77.5 | 73.38 | 72.7 | 74.9 | 69.16 |
| | F1 | 77.2 | 77.6 | 74.98 | 75.13 | 76.7 | 71.9 | 75.5 | **78.9** | 75.06 | 75.2 | 76.5 | 71.04 |
| LR | P | 84 | 79.9 | 79.8 | 82.1 | 74 | 81.4 | 87.8 | 76 | 80.9 | 81.3 | 68.7 | 78 |
| | R | 84 | 79.8 | 79.9 | 81.8 | 74 | 81.2 | 87.7 | 76 | 79.9 | 80.5 | 68.2 | 77.3 |
| | F1 | 84 | 79.9 | 79.9 | 81.5 | 74 | 81.2 | **87.8** | 76 | 80.1 | 80.2 | 68.3 | 77.5 |
| SVM | P | 80.8 | 76.3 | 79.2 | 78.5 | 79.9 | 82.6 | 85.7 | 81.6 | 86.1 | 80.6 | 78.1 | 76.7 |
| | R | 79.9 | 76 | 79.2 | 78.6 | 79.9 | 82.5 | 85.7 | 81.2 | 85.7 | 80.5 | 77.9 | 76 |
| | F1 | 80.1 | 76.1 | 79.2 | 78.4 | 79.9 | 82.4 | **85.7** | 81 | **85.7** | 80.3 | 77.7 | 76 |
| MLP | P | 83.6 | 78.64 | 78.37 | 78.7 | 76.91 | 76.21 | 80.4 | 78.22 | 79.01 | 86.4 | 77.14 | 77 |
| | R | 83.1 | 78.57 | 77.92 | 78.6 | 75.65 | 75.97 | 78.6 | 78.25 | 75.32 | 86.4 | 77.27 | 76.62 |
| | F1 | 83.2 | 78.6 | 78.1 | 78.6 | 76.06 | 76.09 | 78.9 | 78.23 | 76.06 | **86.4** | 77.15 | 76.79 |
| RF | P | 81.16 | 80.44 | 79.21 | 74.96 | 74.44 | 78.31 | 77.45 | 75.35 | 77.43 | 79.39 | 73.79 | 75.92 |
| | R | 77.6 | 79.55 | 78.57 | 75 | 74.03 | 78.25 | 77.27 | 74.68 | 75.65 | 79.22 | 73.38 | 75 |
| | F1 | 78.27 | **79.69** | 78.73 | 74.96 | 74.12 | 78.27 | 77.33 | 74.81 | 76.04 | 79.28 | 73.47 | 75.23 |
| XGB | P | 74.99 | 78.2 | 72.08 | 73.26 | 79.53 | 73.34 | 79.77 | 77.39 | 75.82 | 76.59 | 76.66 | 74.05 |
| | R | 74.03 | 76.95 | 70.78 | 71.75 | 76.95 | 73.05 | 77.6 | 75 | 75 | 75.97 | 75 | 73.05 |
| | F1 | 74.26 | 77.23 | 71 | 72.11 | 77.44 | 73.11 | **78** | 75.46 | 75.13 | 76.13 | 75.35 | 73.21 |

TABLE VIII. PRECISION (P), RECALL (R), AND F1-SCORE (F1) CLASSIFICATION RESULTS (AUTOMATIC ANNOTATED DATASET)

| Feature | | Word Count | | | TF-IDF (word-level) | | | TF-IDF (n-gram-level) | | | TF-IDF (character-level) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Classifier | Measure | Raw Text | Stemming | Rooting | Raw Text | Stemming | Rooting | Raw Text | Stemming | Rooting | Raw Text | Stemming | Rooting |
| NB | P | 76.7 | 75.6 | 72.6 | 76.2 | 76 | 74 | 74.4 | 75.9 | 75.1 | 72.8 | 73.6 | 70.1 |
| | R | 76.4 | 75.5 | 72.5 | 76.2 | 74.8 | 69.4 | 74.4 | 76 | 75 | 72.8 | 73.6 | 70 |
| | F1 | **76.6** | 75.4 | 72.4 | 76.2 | 75.1 | 70.5 | 74.4 | 75.9 | 74.9 | 72.8 | 73.6 | 70 |
| LR | P | 93.4 | 84.6 | 76.3 | 91.8 | 85.8 | 77.1 | 90.8 | 85.8 | 79.7 | 84.3 | 83.1 | 76.9 |
| | R | 93.3 | 84.6 | 76 | 91.7 | 85.8 | 77 | 90.7 | 85.7 | 79.6 | 84.2 | 83.1 | 76.9 |
| | F1 | **93.3** | 84.6 | 76.1 | 91.7 | 85.8 | 77.1 | 90.7 | 85.7 | 79.6 | 84.3 | 83.1 | 76.9 |
| SVM | P | 92.1 | 82.9 | 78.3 | 91.2 | 84.2 | 79.8 | 90.6 | 85.4 | 82 | 89.7 | 83.8 | 76 |
| | R | 92 | 82.8 | 78 | 91.2 | 84.2 | 79.6 | 90.5 | 85.3 | 81.8 | 89.1 | 83.1 | 75.1 |
| | F1 | **92** | 82.8 | 78 | 91.2 | 84.2 | 79.7 | 90.5 | 85.3 | 81.9 | 89.4 | 83.3 | 75.4 |
| MLP | P | 88.8 | 79.6 | 70.1 | 87.1 | 77 | 70.4 | 71.7 | 73.2 | 72.2 | 80.5 | 78 | 72.3 |
| | R | 88.5 | 79.5 | 70 | 87.1 | 77 | 70.4 | 71.7 | 73.2 | 72.2 | 80.5 | 78 | 72.3 |
| | F1 | **88.6** | 79.5 | 70.1 | 87.1 | 77 | 70.4 | 71.7 | 73.2 | 72.2 | 80.5 | 78 | 72.3 |
| RF | P | 84.9 | 82.5 | 77.7 | 84.6 | 82.1 | 77.7 | 84.9 | 81.5 | 78.3 | 78.3 | 79 | 76.1 |
| | R | 84.7 | 82.3 | 77.5 | 84.7 | 82.1 | 77.6 | 84.9 | 81.6 | 78.3 | 78.3 | 78.9 | 76.1 |
| | F1 | 84.7 | 82.4 | 77.6 | 84.6 | 82.1 | 77.6 | **84.9** | 81.5 | 78.3 | 78.3 | 78.9 | 76.1 |
| XGB | P | 82.8 | 82.1 | 76.2 | 82.8 | 81.7 | 75.8 | 82.3 | 82 | 76.3 | 80.1 | 80.1 | 76.5 |
| | R | 80.2 | 80.4 | 74.7 | 80.9 | 80.5 | 75.2 | 79.9 | 80.4 | 75.2 | 79.3 | 79.3 | 75.8 |
| | F1 | **80.7** | **80.7** | 75.1 | 81.2 | **80.7** | 75.3 | 80.4 | 80.7 | 75.5 | 79.5 | 79.5 | 76 |

enhance the classification results with the text from social media.

As shown in Fig. 4, the highest recall was obtained using the count vector TF-IDF feature with the LR classifier (87.7%), and the count vector feature with the LR classifier (93.3%) on manually and automatically annotated corpora, respectively. The highest F1-score, as shown in Fig. 5 was obtained using the n-gram-level TF-IDF feature with the MLP classifier (87.8%) and the count vector feature with the LR classifier (93.3%) on manually and automatically annotated corpora, respectively.

## VI. DISCUSSION

The primary objective of this research was to build a benchmark dataset for fake news in Arabic related to the COVID-19 pandemic. We introduce a new fake news corpus in the Arabic language, collected from Twitter. It is clear from the experimental results that the manually annotated corpus can be used as a baseline for further research in the domain of fake news and misinformation. As there remains no benchmark dataset for fake news detection in Arabic related to the COVID-19 pandemic, this corpus will help the research community once the dataset is publicly available. The proposed corpus was manually annotated by three annotators to ensure the quality and usefulness of the developed corpus. We used a set of machine learning classifiers to train different machine learning models on the manually annotated corpus. The best model was selected to predict fake news classes of unlabeled tweets (more than 35,000 tweets). The statistical analysis showed lower precision, recall, and F1-score values in the classification of the manually annotated corpus, while the automatically annotated corpus showed improved results. From the results presented in the previous section, we notice that increasing the size of the dataset leads to an improvement in the classification results using precision, recall, and F1-score measures.

The use of machine learning methods to classify the fake news corpus using content-based features gives better results than the user-based features. The corpus can be further expanded using two methods: 1) increasing the number of

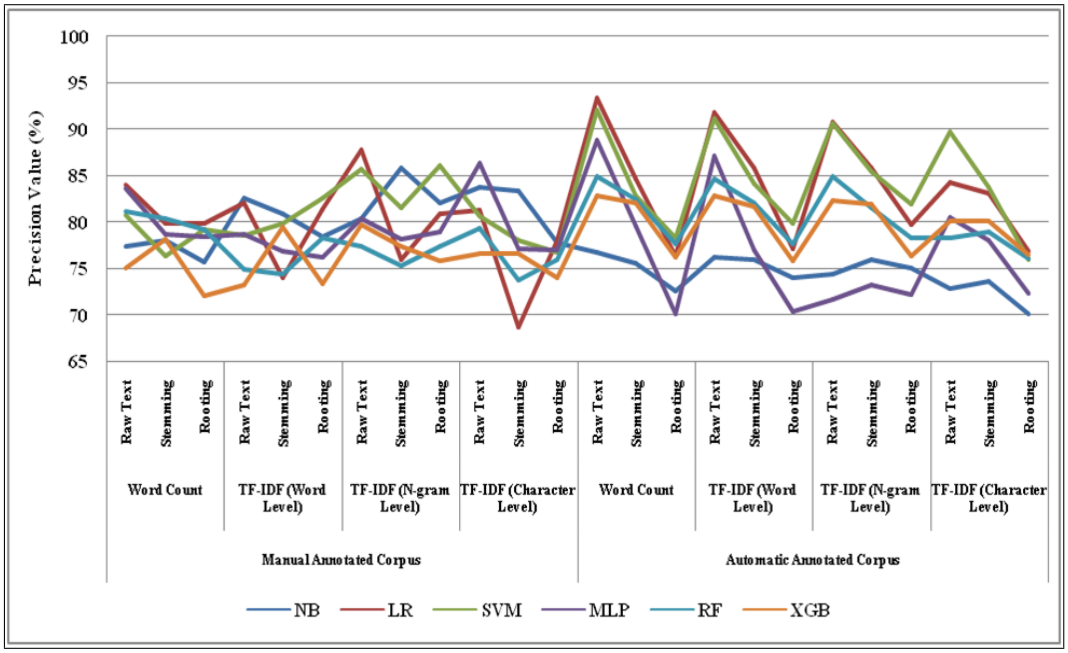


Fig. 3. Precision Results.

verified rumouror misinformation topics, or 2) performing classification on more unlabeled tweets related to the COVID-19 pandemic. After then, the deep learning approach can be used to enhance fake news classification.

## VII. Conclusion and Future Work

In this paper, we introduced a new Arabic corpus of fake news that will be made publicly available for research purposes on this link: (https://github.com/yemen2016/FakeNewsDetection), after preparing the tweets ID's and their associated classes. We explained the collection process of fake news and gave details about how we select rumors and misinformation topics during the COVID-19 pandemic. The classification task was performed using six classifiers (Naïve Bayes, Logistic Regression, Support Vector Machine, Multilayer Perceptron, Random Forest Bagging, and eXtreme Gradient Boosting) to test the possibility of recognizing fake and genuine tweets. We used four feature types: count vector, word-level TF-IDF, n-gram-level TF-IDF, and character-level TF-IDF. We noticed that the achieved performance varies depending on the features and classifiers used. Along with considering the raw text as an input to the machine learning classifiers, we also used two pre-processing methods: stemming and rooting. Both techniques failed to improve the classification results as the corpus text was collected from Twitter, which includes various dialects and language mistakes. Therefore, the stemming and rooting procedures did not produce correct results. The study concluded that we can achieve higher performance with more annotated data.

In the future, we plan to expand our corpus with additional verified rumour and misinformation topics. We also look forward to investigating the performance of new classification methods such as deep learning. In this research, we only used content-based features to classify and analyze fake news, though user-based features may also be utilized.

## Acknowledgment

"This work was supported by the research Project PSU-COVID19 Emergency Research Program; Prince Sultan University; Saudi Arabia [PSU-COVID19 Emergency Research Program-CCIS-2020-57]".

## References

[1] J. AlHumaid, S. Ali, and I. Farooq, "The psychological effects of the covid-19 pandemic and coping with them in saudi arabia." *Psychological Trauma: Theory, Research, Practice, and Policy*, vol. 12, no. 5, p. 505, 2020.

[2] H. B. Dunn and C. A. Allen, "Rumors, urban legends and internet hoaxes," in *Proceedings of the Annual Meeting of the Association of Collegiate Marketing Educators*, 2005, p. 85.

[3] N. DiFonzo and P. Bordia, "Rumor, gossip and urban legends," *Diogenes*, vol. 54, no. 1, pp. 19–35, 2007.

[4] M. K. Elhadad, K. F. Li, and F. Gebali, "Covid-19-fakes: a twitter (arabic/english) dataset for detecting misleading information on covid-19," in *International Conference on Intelligent Networking and Collaborative Systems*. Springer, 2020, pp. 256–268.

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

[6] M. K. Elhadad, K. F. Li, and F. Gebali, "Fake news detection on social media: a systematic survey," in *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*. IEEE, 2019, pp. 1–8.

[7] P. L. Liu and L. V. Huang, "Digital disinformation about covid-19 and the third-person effect: examining the channel differences and negative emotional outcomes," *Cyberpsychology, Behavior, and Social Networking*, vol. 23, no. 11, pp. 789–793, 2020.

[8] N. M. Krause, I. Freiling, B. Beets, and D. Brossard, "Fact-checking as risk communication: the multi-layered risk of misinformation in times of covid-19," *Journal of Risk Research*, vol. 23, no. 7-8, pp. 1052–1059, 2020.

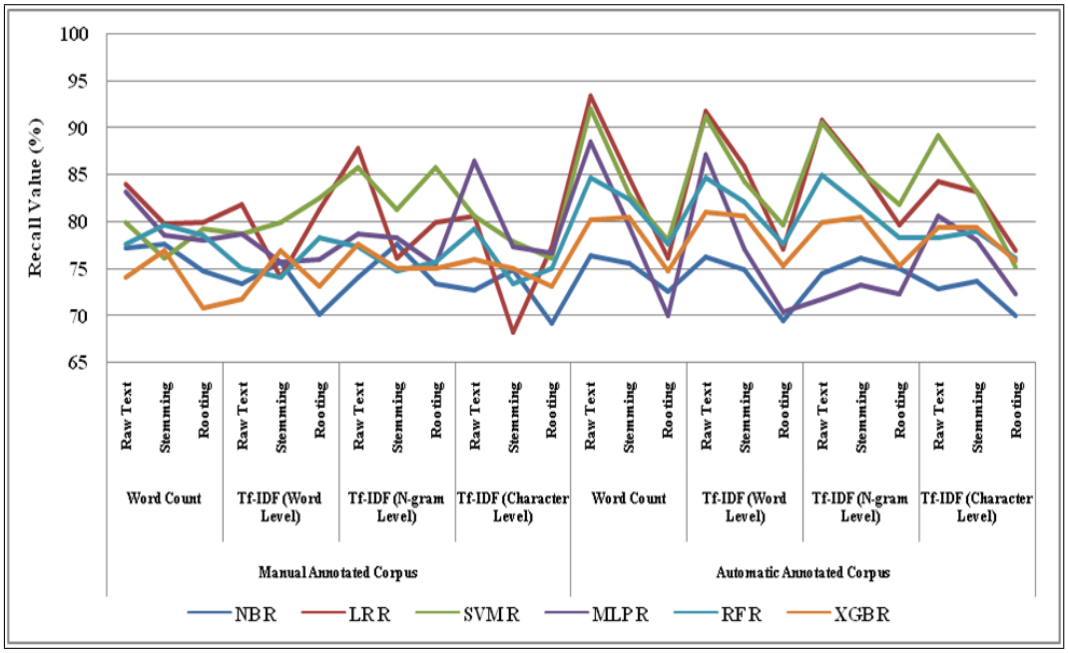


Fig. 4. Recall Results.

[9] J. Donovan, "Concrete recommendations for cutting through misinformation during the covid-19 pandemic," 2020.

[10] M. Luengo and D. García-Marín, "The performance of truth: politicians, fact-checking journalism, and the struggle to tackle covid-19 misinformation," *American Journal of Cultural Sociology*, vol. 8, no. 3, pp. 405–427, 2020.

[11] F. Alam, F. Dalvi, S. Shaar, N. Durrani, H. Mubarak, A. Nikolov, G. D. S. Martino, A. Abdelali, H. Sajjad, K. Darwish *et al.*, "Fighting the covid-19 infodemic in social media: a holistic perspective and a call to arms," *arXiv preprint arXiv:2007.07996*, 2020.

[12] L. Alsudias and P. Rayson, "Covid-19 and arabic twitter: How can arab world governments and public health organizations learn from social media?" in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, 2020.

[13] S. A. Alkhodair, S. H. Ding, B. C. Fung, and J. Liu, "Detecting breaking news rumors of emerging topics in social media," *Information Processing & Management*, vol. 57, no. 2, p. 102018, 2020.

[14] S. M. Alzanin and A. M. Azmi, "Rumor detection in arabic tweets using semi-supervised and unsupervised expectation–maximization," *Knowledge-Based Systems*, vol. 185, p. 104945, 2019.

[15] N. Y. Hassan, W. H. Gomaa, G. A. Khoriba, and M. H. Haggag, "Supervised learning approach for twitter credibility detection," in *2018 13th International Conference on Computer Engineering and Systems (ICCES)*. IEEE, 2018, pp. 196–201.

[16] A. Habib, S. Akbar, M. Z. Asghar, A. M. Khattak, R. Ali, and U. Batool, "Rumor detection in business reviews using supervised machine learning," in *2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)*. IEEE, 2018, pp. 233–237.

[17] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*. Springer, 2017, pp. 127–138.

[18] M. Alkhair, K. Meftouh, K. Smaïli, and N. Othman, "An arabic corpus of fake news: Collection, analysis and classification," in *International Conference on Arabic Language Processing*. Springer, 2019, pp. 292–302.

[19] F. A. Ozbay and B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms," *Physica A: Statistical Mechanics and its Applications*, vol. 540, p. 123174, 2020.

[20] F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "Arcov-19: The first arabic covid-19 twitter dataset with propagation networks," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 82–91.

[21] S. Alqurashi, A. Alashaikh, and E. Alanazi, "Identifying information superspreaders of covid-19 from arabic tweets," *Preprints*, 2020.

[22] G. K. Shahi and D. Nandini, "Fakecovid–a multilingual cross-domain fact check news dataset for covid-19," *arXiv preprint arXiv:2006.11343*, 2020.

[23] D. Dimitrov, E. Baran, P. Fafalios, R. Yu, X. Zhu, M. Zloch, and S. Dietze, "Tweetscov19-a knowledge base of semantically annotated tweets about the covid-19 pandemic," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2991–2998.

[24] S. A. Memon and K. M. Carley, "Characterizing covid-19 misinformation communities using a novel twitter dataset," *arXiv preprint arXiv:2008.00791*, 2020.

[25] Y. Li, B. Jiang, K. Shu, and H. Liu, "Mm-covid: A multilingual and multidimensional data repository for combatingcovid-19 fake new," *arXiv preprint arXiv:2011.04088*, 2020.

[26] L. Cui and D. Lee, "Coaid: Covid-19 healthcare misinformation dataset," *arXiv preprint arXiv:2006.00885*, 2020.

[27] O. Oueslati, E. Cambria, M. B. HajHmida, and H. Ounelli, "A review of sentiment analysis research in arabic language," *Future Generation Computer Systems*, vol. 112, pp. 408–430, 2020.

[28] L. Rosenzweig, B. Bago, A. J. Berinsky, and D. Rand, "Misinformation and emotions in nigeria: The case of covid-19 fake news," 2020.

[29] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective computing and sentiment analysis," in *A practical guide to sentiment analysis*. Springer, 2017, pp. 1–10.

[30] A. Al-Laith and M. Shahbaz, "Tracking sentiment towards news entities from arabic news on social media," *Future Generation Computer Systems*, vol. 118, pp. 467–484, 2021.

[31] A. Al-Laith, M. Shahbaz, H. F. Alaskar, and A. Rehmat, "Arasencorpus: A semi-supervised approach for sentiment annotation of a large arabic text corpus," *Applied Sciences*, vol. 11, no. 5, p. 2434, 2021.

[32] A. Al-Laith and M. Alenezi, "Monitoring people's emotions and symptoms from arabic tweets during the covid-19 pandemic," *Information*, vol. 12, no. 2, p. 86, 2021.

[33] M. Z. Ali, Ehsan-Ul-Haq, S. Rauf, K. Javed, and S. Hussain, "Improving

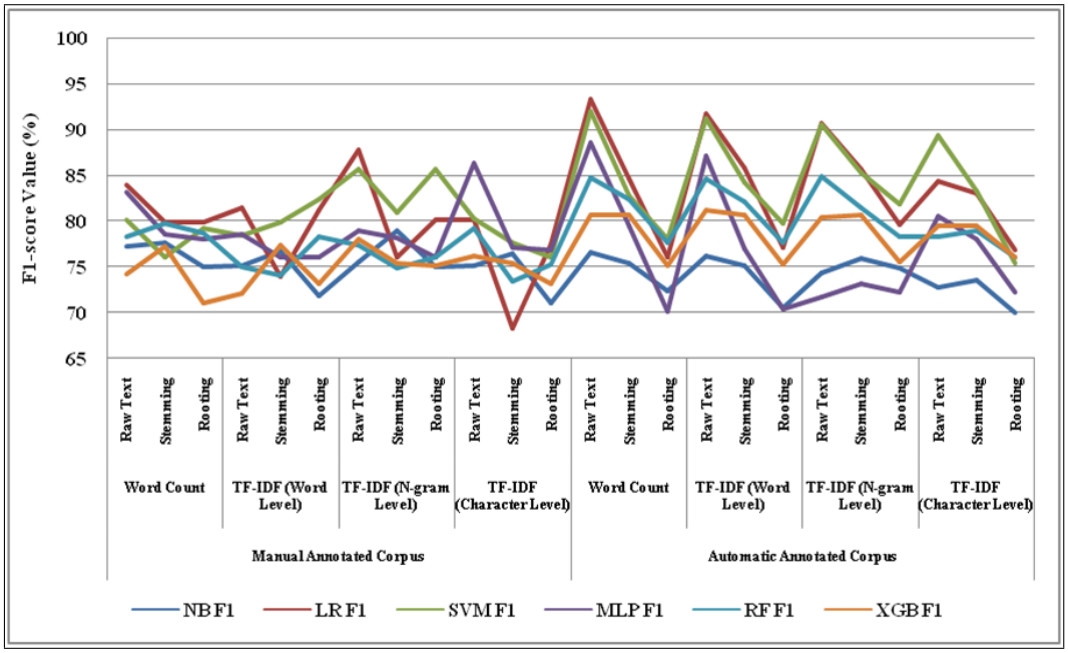


Fig. 5. F1-score Results.

hate speech detection of urdu tweets using sentiment analysis," *IEEE Access*, vol. 9, pp. 84 296–84 305, 2021.

[34] A. Allaith, M. Shahbaz, and M. Alkoli, "Neural network approach for irony detection from arabic text on social media." in *FIRE (Working Notes)*, 2019, pp. 445–450.

[35] S. R. El-Beltagy, M. E. Kalamawy, and A. B. Soliman, "Niletmrg at semeval-2017 task 4: Arabic sentiment analysis," *arXiv preprint arXiv:1710.08458*, 2017.