DISINFORMATION DETECTION ABOUT ISLAMIC ISSUES ON SOCIAL MEDIA USING DEEP LEARNING TECHNIQUES

Suhaib Kh. Hamed[1],* and Mohd Juzaiddin Ab Aziz[1] and Mohd Ridzwan Yaakub[2]
[1] Center for Software Technology and Management (SOFTAM), Faculty of Information Science and Technology, University Kebangsaan Malaysia (UKM), Bangi 43600, Selangor, Malaysia; juzaiddin@ukm.edu.my
[2] Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, University Kebangsaan Malaysia (UKM), Bangi 43600, Selangor, Malaysia; ridzwanyaakub@ukm.edu.my
* Correspondence: p105401@siswa.ukm.edu.my

ABSTRACT

Nowadays, many people receive news and information about what is happening around them from social media networks. These social media platforms are available free of charge and allow anyone to post news or information or express their opinion without any restrictions or verification, thus contributing to the dissemination of disinformation. Recently, disinformation about Islam has spread through pages and groups on social media dedicated to attacking the Islamic religion. Many studies have provided models for detecting fake news or misleading information in many domains, such as political, social, economic, and medical, except in the Islamic domain. Due to this negative impact of spreading disinformation targeting the Islamic religion, there is an increase in Islamophobia, which threatens societal peace. In this paper, we present a Bidirectional Long Short-Term Memory-based model trained on an Islamic dataset (RIDI) that was collected and labeled by two separate specialized groups. In addition, using a pre-trained word-embedding model will generate Out-Of-Vocabulary, because it deals with a specific domain. To address this issue, we have retrained the pre-trained Global Vectors for Word Representation (GloVe) model on Islamic documents using the Mittens method. The results of the experiments proved that our proposed model based on Bidirectional Long Short-Term Memory with the retrained GloVe model on the Islamic articles is efficient in dealing with text sequences better than unidirectional models and provides a detection accuracy of 95.42% of Area under the ROC Curve measure compared to the other models.

KEYWORDS: Disinformation Detection, Fake News, Social Media, Deep Learning, Islamic Domain

1.0 INTRODUCTION

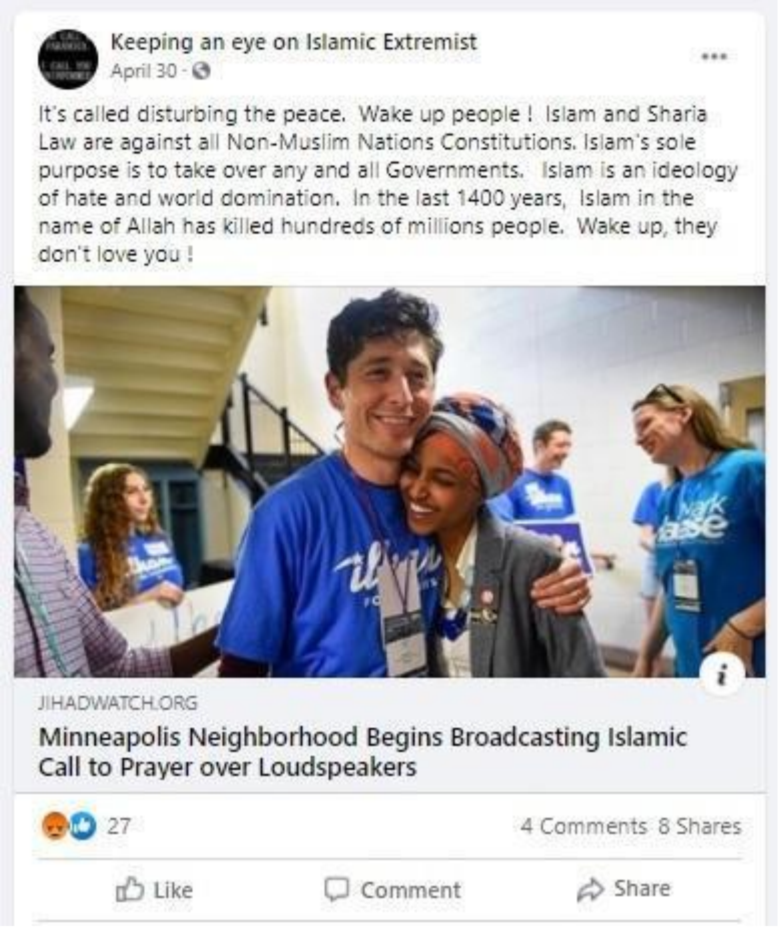
Nowadays, social media has become the main source for receiving news and information [1]. Unlike traditional media, anyone can post news or information on social media at any time, free of cost and without any restrictions, leading to a massive spread of false information [2]. Social media platforms have provided a virtual environment for discussion, exchange of views, and global interaction between users [3], without restrictions on location, time, or content volume [4]. The spread of false information on social media has become a global threat [5], and it may create negative impacts on societal peace [6] and affect people's daily lives [7]. False information is divided into two main groups: misinformation and disinformation. Misinformation is false information that is shared unintentionally and without any bad intent, while disinformation is false information that is intentionally shared and spread with disruptive intent [8]. The manipulation of facts or the dissemination of outdated, incorrect, and unconfirmed information to perplex the audience and sway their opinion is referred to as "disinformation." Fake news, on the other hand, is often thought of as untrue information, about specific events, that primarily exists online [9]. The propagation of disinformation that targets the Islamic religion leads to the growth of hate speech against Muslims [10], which may cause damage to the reputation of Muslims and their interests and may also lead to a rift in the peaceful coexistence of society, of which Muslims are one of the components [11]. The field of fake news detection in social media has garnered considerable interest from researchers and is considered one of the emerging fields [5]; it needs more development [12]. One of the common challenges is that false information data is diverse in terms of subjects and styles, and is relevant to specific events [13]. The dearth of models to curb the spread of disinformation or fake news on Islamic issues on social media helps the emergence of anti-Islam movements and increases Islamophobia [14], [10]. The existing models are trained on existing standard datasets; most of these datasets consist of political, economic, or social news [15] and cannot detect disinformation or fake news in all domains, and thus present poor results in specific or other domains [16], [17], [18], [5], [13], [19], [20]. One of the problems faced by NLP-based models targeting specific domains is Out-Of-Vocabulary (OOV) words. Pre-trained Word2Vec model cannot handle OOV words [21]. The FastText model can handle OOV by providing embeddings for characters based on N-grams. The Fasttext model's large memory requirements are a major shortcoming since it embeds words based on their characters [22]. In this research, pre-trained GloVe model will be used to define OOV words, then retrain GloVe model on the dataset used to show the importance of terms related to the specific domain in improving detection. Artificial intelligence (AI) and Natural Language Processing (NLP)-related problems have seen extensive use of Deep Learning (DL) approaches during the past ten years [23]. DL approaches are currently attracting researchers' attention and have become quite well-liked in the AI research community in recent years [24], [25]. This is due to findings that traditional Machine Learning (ML) techniques cannot match them in a wide range of sectors [26], [4], [6]. Therefore, it is necessary to develop a new model to detect disinformation that targets Islamic issues by using a deep learning technique that has proven its accuracy in detecting fake news [27]. The contributions of this research could be summarized as follows:

- Providing a standard dataset related to the disinformation detection about Islamic issues called "CIDII"
- Examining the importance of features related to writing styles such as exclamation marks, question marks, and capital letters in detecting fake news
- Presenting word embedding model based on GloVe for the Islamic domain.
- Proposing a model based on Bi-LSTM and retrained GloVe on Islamic documents to detect disinformation on Islamic issues using a real dataset.

The remainder of this paper is organized as follows: Section 2.0 highlights the problem background of the research, section 3.0 presents the previous studies, provides an overview of the features used in fake news-detection models, and reviews the deep-learning techniques used to detect fake news, as well as the related works. Section 4.0 explains the methodology and dataset used in this research, while section 5.0 provides the obtained results and discusses these findings. In section 6.0, the conclusion of the paper is presented.

2.0 SPREAD OF DISINFORMATION ABOUT ISLAM

Disinformation related to religious matters is one of the critical issues, because it may remain for long periods, unlike rumors or fake news related to public issues, which will disappear after some time or be refuted. In addition, in case of that such deceptive information is not discovered or is not prevented from spreading, such disinformation may remain in the reader's mind [2]. The falsification of facts and the dissemination of false information concerning the Islamic religion as a material, and the misrepresentation of the history of its symbols, raises many interactions with this disinformation by sharing or commenting on it, and increases hate speech and thus creates a gap between people of the same community (between Muslims on the one hand and other religions on the other), especially in countries where Muslims are considered one of the components of society or within the minorities. These speeches aim to highlight them as backward or terrorists from some extremist individuals or groups that incite hatred, which encourages their targeting or creates a gap in society against peaceful coexistence [28], [2]. Many individuals, particularly from outside the Islamic religion, who have not been familiar with the teachings and rulings of Islam and have no preconceived idea of them, will believe these lies, especially those who derive their information from pages or personal accounts on social media. Islamophobia is one of the consequences of spreading disinformation about Islam. Islamophobia is a phenomenon characterized by showing hostility or hatred against people or institutions because of their affiliation with the Islamic religion through the practice of several methods, including violence, exclusion, or discrimination. An actual example of Islamophobia is harassment of Muslims because of their appearance and clothing, which may be associated with verbal or physical assault, discrimination against them in public institutions, or attacks on their mosques [11]. Some reports from Scotland Yard as well as from some organizations in the United Kingdom (UK) such as Tell Measuring Anti-Muslim Attacks (MAMA) indicated an increase in violations and hate speeches against Muslims online, as the report included the creation of many Facebook pages to target Muslims in the UK [29]. Among these examples of disinformation targeting Islam are Figures 1 and 2. As Figure. 2 represents the disinformation about the concept of "Jizya" posted on one of the anti-Islam Facebook pages, while Figure 3 represents the correct information about the concept of "Jizya" published on Wikipedia (https://en.wikipedia.org/wiki/Jizya, accessed on 16 November 2020). Most of the previous studies did not employ the techniques in how to know the motives behind the spread of fake news and contented themselves with providing models for detecting fake news. The motivations and reasons for this destructive behavior are mostly unknown and under-researched [30]. Theoretical frameworks for comprehending the dissemination of false information are lacking (intentionally or unintentionally) [31]. In general, fake news comes in a variety of formats and motivations, such as clickbait which seeks to generate revenue through clicks, and politically driven fake news which tries to bring down candidates in the elections [32]. In the Islamic domain, there are several motives behind spreading disinformation about Islam on some social media pages, one of these motives is that these pages belong to extreme right-wing parties that may benefit from inflaming attitudes against Muslims or immigrants to gain the votes of the other party and mobilize them in their favor. As an example, Facebook pages were used by a Danish right-wing party to spread hate and fake news against Muslims [28]. The other motive is that these pages may belong to groups claiming to be affiliated with Islam, but they publish deceptive content by spreading distorted Qur'anic verses, wrong interpretations, or using negative emotions, to show a bad impression of Islam [33]. In addition, among these motives, some of these pages belong to extremist groups that carry extremist ideologies that contradict the tolerant teachings of Islam or adopt ideas stemming from a wrong interpretation of the rulings of Islam, to brainwash people who do not have prior knowledge of the Islamic religion and its noble principles to attract them to implement their goals by claiming that they are applying God's law [28]. However, there are some pages on social media run by Muslims that unintentionally publish Misinformation that contribute to distorting the image of the Islamic religion by promoting false scientific information, displaying illogical statistical numbers, or excessive generalizations based on unreliable and unverified resources [14]. As for the Islamic domain, there is a dearth of previous studies that discussed fake news on social media that relate to or targets the religious aspect in general or the Islamic aspect in particular, despite their seriousness and negative impacts on society. Vidgen, Yasseri [10] used a Support Vector Machine (SVM) classifier to identify strong and weak Islamophobia hate speech on social media. They said that creating a dataset containing sufficient examples of hate speech for training is a tedious process. Therefore, to address this issue, they sampled relevant data from Twitter, selecting 1000 tweets from a total of 4000 tweets related to Islamophobia hate speech using the two search words (Islam, Muslims) under the supervision of domain experts. In order to provide the best performance of the classifier and the possibility of providing high results when applied in the real-world environment, they proceeded to expand the state of heterogeneity in the selected instances and include various examples of linguistic patterns. They mentioned to create a balanced dataset; the number of anti-Islamic hate speech examples was reduced so that all categories were balanced. Regarding the word embeddings, they created a GloVe model that was trained on a set of tweets related to their field of research, where they pointed out that training a word embeddings model for a specific domain presents promising results. They stated that the domain of the word embeddings is more important than the size of the word embeddings. Therefore, there is an urgent need to introduce efficient detection models using deep learning techniques to combat the spread of this disinformation.



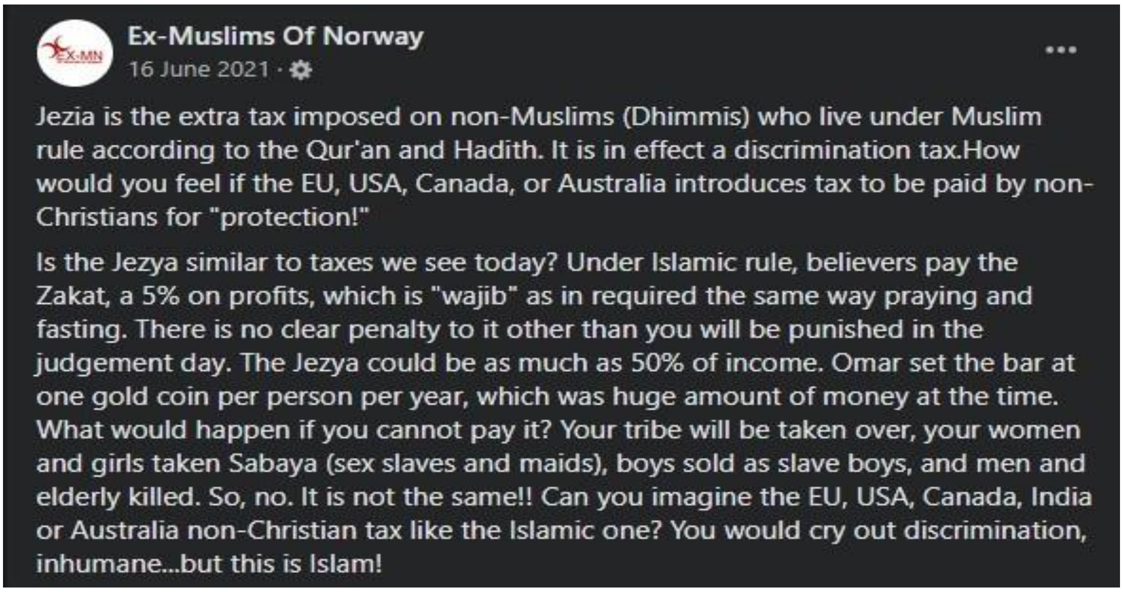
Keeping an eye on Islamic Extremist
April 30 ·

It's called disturbing the peace. Wake up people ! Islam and Sharia Law are against all Non-Muslim Nations Constitutions. Islam's sole purpose is to take over any and all Governments. Islam is an ideology of hate and world domination. In the last 1400 years, Islam in the name of Allah has killed hundreds of millions people. Wake up, they don't love you !

JIHADWATCH.ORG
Minneapolis Neighborhood Begins Broadcasting Islamic Call to Prayer over Loudspeakers

27 4 Comments 8 Shares

Fig. 1: The disinformation post on one of the Facebook pages (https://www.facebook.com/everhartsgarage, accessed on 21 November 2020).

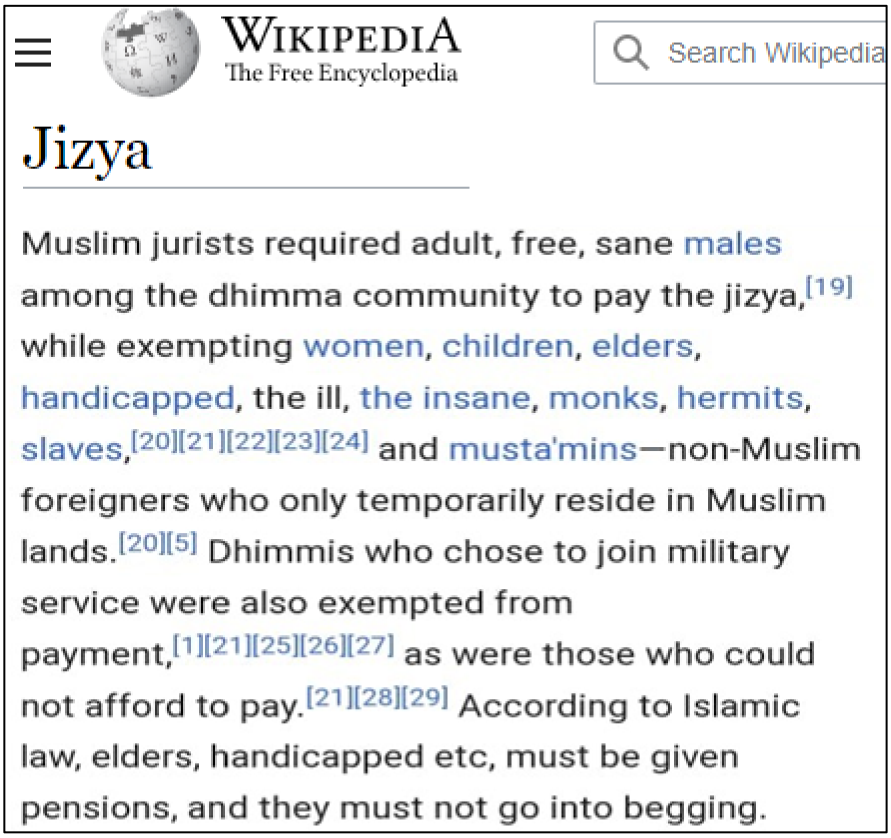


Ex-Muslims Of Norway
16 June 2021 ·

Jezia is the extra tax imposed on non-Muslims (Dhimmis) who live under Muslim rule according to the Qur'an and Hadith. It is in effect a discrimination tax.How would you feel if the EU, USA, Canada, or Australia introduces tax to be paid by non-Christians for "protection!"

Is the Jezya similar to taxes we see today? Under Islamic rule, believers pay the Zakat, a 5% on profits, which is "wajib" as in required the same way praying and fasting. There is no clear penalty to it other than you will be punished in the judgement day. The Jezya could be as much as 50% of income. Omar set the bar at one gold coin per person per year, which was huge amount of money at the time. What would happen if you cannot pay it? Your tribe will be taken over, your women and girls taken Sabaya (sex slaves and maids), boys sold as slave boys, and men and elderly killed. So, no. It is not the same!! Can you imagine the EU, USA, Canada, India or Australia non-Christian tax like the Islamic one? You would cry out discrimination, inhumane...but this is Islam!

Fig. 2: The disinformation post on one of the Facebook pages (https://www.facebook.com/exmuslims.no/posts/pfbid02EcTXBjcTjnq4dxE12a8n4brCVikFe6iW3YtDUcVXyyVx4zJNirHdWfCXu5qgYVRbl/, accessed on 26 March 2022).



Jizya

Muslim jurists required adult, free, sane males among the dhimma community to pay the jizya,[19] while exempting women, children, elders, handicapped, the ill, the insane, monks, hermits, slaves,[20][21][22][23][24] and musta'mins—non-Muslim foreigners who only temporarily reside in Muslim lands.[20][5] Dhimmis who chose to join military service were also exempted from payment,[1][21][25][26][27] as were those who could not afford to pay.[21][28][29] According to Islamic law, elders, handicapped etc, must be given pensions, and they must not go into begging.

Fig. 3: An example of the correct information regarding the concept of “Jizya” according to the perspective of the Islamic religion published on the Wikipedia page, (This paragraph as a screenshot is excerpted from the Wikipedia article on the Jizya).

3.0 PREVIOUS STUDIES

This section is divided into three subsections. Section 3.1 examines the use of content-based features in detecting fake news. Section 3.2 reviews studies that provided detection models using the DL methods. In addition, section 3.3 discusses relevant studies.

3.1. Content-Based Fake News Detection Models

Content-based deep learning models in detecting fake news mean that they rely mainly on features that a classifier can extract from content [34], such as linguistic features, syntactic features [25], or sentiment-based features [35]. The most prevalent linguistic features are, 1) lexical features, such as character-level and word-level features like total words, characters per word, frequency of large words, and unique words; and 2) syntactic features, such as sentence-level traits like frequency of function words and phrases, or punctuation [36]. Moreover, domain-specific language features, such as entity names, terminology, and quoted words, are specifically compatible with the news or information domain. Based on studies and experiments showing that unreliable content tends to be more emotional than reliable information, sentiment or emotion polarity cues can be a key indicator to identify reliable from unreliable content [37], and mostly fake news or disinformation tends to have aggressive emotions [38]. Another important feature in detecting disinformation is the feature based on the style and quality of writing [39]. Fake news is often poorly worded and contains slang, misspelled, profanity, or repeated characters unusually, or it contains more question and exclamation marks than real news [40], or writing words in capital letters [41]. Moreover, the vocabulary is used frequently in fake news, unlike fake news which contains a variety of vocabulary [42]. In addition to the features based on article size such as fake news headlines are often longer than real news headlines, while fake news text is shorter than real news text [43]. The existing content-based detection models can be divided into two types, Knowledge-based and Style-based. 1) Knowledge-based methods leverage external resources to fact-check the claims in the content of information. A claim's truthiness value in a given context is intended to be determined by fact-checking. 2) Style-based: disinformation publishers frequently have the malicious intent to disseminate disinformation and influence large communities of consumers, necessitating specific writing styles that are required to appeal to and persuade a wide scope of readers and are not seen in real information articles. Style-based methods aim to identify manipulators in the writing of article content in order to identify disinformation [36].

3.2. Fake News Detection Based On Deep Learning Techniques

Deep Learning (DL) is a technique that uses Neural Networks (NN) to imitate the mechanisms of the human brain to find patterns [44]. The basic structure of Neural Networks consists of three types of layers: Input layer, hidden layer(s), and output layer. These layers consist of neurons connected, which are associated with randomly initialized weights. Neurons receive input values through the input layer, and these NNs use an activation function on the data to standardize the outputs resulting from neurons. The process of iteration during the stage of training the NN on the labeled dataset leads to the production of outputs, and these outputs are compared with the resulting cost function, which indicates the extent to which the prediction of the algorithm deviates from the real outputs, and accordingly, the weights associated with the neurons will be adjusted through the use of the optimization function after each iteration, to reduce the deviation of the cost function [44]. Deep learning classifiers have gained popularity recently and are effective in extracting relevant features [45]. DL technique automatically learns features from big data as opposed to using hand-crafted features, which is the main distinction between DL and conventional data classification techniques [46]. Among these DL techniques that have proven their efficiency in this field is the Recurrent Neural Networks (RNN) model and its variants, such as Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and Bidirectional LSTM (Bi-LSTM) models [47], because it is already able to capture the full meaning of a text and the context of an article from the flow of its input string [48]. A deep learning model like Bi-LSTM is a successful classifier when dealing with sequential data, like text or time series. It can perform better than other classifiers since it respects the word order [49]. The Bi-LSTM has proven its efficiency in detecting fake news and it is better than unidirectional approaches [50]. Kumar, Asthana [41] concluded that the BI-LSTM classifier greatly outperformed simple Convolutional Neural Networks (CNN) architecture in fake news detection. Asghar, Habib [51] mentioned in their research that the Bi-LSTM model demonstrated effective performance in rumor detection by using the CNN layer for extracting valuable features.

3.3. Related Works Based On A Specific Domain

Many studies have presented models to detect fake news that targets a general domain i.e. the datasets used in these models contain news on a variety of topics (political, social, economic news, etc.), and also there are a large number of fake news detection studies targeting the domain of politics [15]. However, some studies targeted specific domains other than political or social in recent years. In one of the studies related to the technology domain Barbado, Araque [52] develops a feature structure based on review text and user behavior to identify fake reviews in the consumer electronics domain. In four separate cities, they created a dataset to detect fake reviews in the consumer electronics industry. According to their study, user-specific features have been proven to be more successful. The ones resulting from the reviewing process are the most pertinent. Their experiments showed high classification accuracy when they used Ada Boost classifier to detect fake reviews. Regarding studies that have investigated a specific domain, including studies that aim to detect misinformation about Covid-19, Elhadad, Li [53], suggested a detection model of misinformation on COVID-19 based on the voting ensemble machine learning classifier that depends on the World Health Organization (WHO), the United Nations Children's Fund (UNICEF), and the United Nations (UN) as sources of information together with epidemiological data gathered from a variety of fact-checking websites. The evaluation's findings show how well the ground truth data were gathered, how valid they were, and how well they worked to build models to detect misinformation. To detect misleading information on COVID-19, Abdelminaam, Ismail [45] proposed an effective and improved deep learning technique based on modified-LSTM (2 layers). They test their methodology on three additional datasets in addition to their experiment on a sizable COVID-19-related dataset (disasters, PolitiFact, and gossip cop). The results show that the proposed framework is highly accurate at detecting misinformation tweets that targeted COVID-19. Tashtoush, Alrababah [20] developed a model based on CNN to automatically classify fake news articles about COVID-19 posted on social media. The "COVID-19 Fake News" dataset was used to train and test their model. The real information of this dataset was collected from the reliable source, which are: the WHO, the International Committee of the Red Cross (ICRC), the UN, the UNICEF, and their official Twitter accounts, where the false information was gathered from various fact-checking websites (such as Snopes, PolitiFact, and FactCheck). Their model's evaluation findings showed great accuracy in identifying COVID-19 misinformation.

4.0 METHODOLOGY

This section consists of three subsections. Section 3.1 highlights the dataset used in this study, and Section 3.2 presents in detail the proposed methodology for providing the fake news detection model. Section 3.3 discusses the methods of evaluating the proposed model.

4.1 Dataset

The most recent labeled and benchmark dataset with reliable ground truth labels is one of the challenges in detecting disinformation and fake news [54]. There are many standard datasets for identifying fake news in several fields, such as the political, social, and economic fields. Recently, datasets related to detecting misinformation regarding COVID-19 were presented in the medical field [55]. These datasets do not contribute to training a model capable of detecting disinformation targeting the Islamic religion, because each dataset has a structure, features, and domain-specific language that differs from the other, and without accurate training data, a detection model is useless. Therefore, regarding the proposed detection model of disinformation related to the Islamic domain, a relevant dataset was collected and used in our research. This dataset was used to identify disinformation about Islamic issues. It is known that the Islamic field is wide and includes several topics related to the Islamic religion, so this research will investigate two topics on which there is much confusion in non-Muslim societies:

- Women in Islam in terms of their rights and duties.
- How does the Islamic religion command Muslims to deal with people of other religions?

Therefore, most of the data collected are related to these two topics. This section includes the following subsections. Section 4.1.1 explains the dataset collection process, and Section 4.1.2 provides the dataset labeling procedure. Section 4.1.3 describes thecollected dataset and Section 4.1.4 shows the visualization of the dataset.

4.1.1. Dataset Collection

The data was collected by a team of three Ph.D. candidates from similar scientific backgrounds specializing in Computer Science and Information Technology, including the researcher in this study. In addition, they have knowledge of the Islamic religion. The posts related to Islamic issues were gathered from Facebook for the period from 1/1/2016 to 20/1/2021, and the process of collecting this data took three months. Only data written in English were collected. The size of the dataset of “Correct Information and Disinformation about Islamic Issues” that was collected is 731 examples as illustrated in Table 1. We called this dataset “CIDII”, which stands for “Correct Information and Disinformation about Islamic Issues”.

Table 1: The size of the CIDII dataset

Class	No. of Examples
Correct Information	431
Disinformation	300

Based on the scope of our research, we collected Correct information about the Islamic religion from four significant Islamic sources [56] as shown in Table 2. Therefore, the writing style and vocabulary used by the sources were considered to cover all possibilities.

Table 2: Correct information sources

Correct Information Source Name	Correct Information Source Link
Noble Qur’an in English	https://quran.com/en
Hadith	https://sunnah.com/
Ibn-Kathir Interpretation	https://www.alim.org/quran/tafsir/ibn-kathir/
Islamic Page	https://islamonline.net/

As for the disinformation related to the Islamic religion, it was collected from pages and groups on Facebook dedicated to spreading against Islam and Muslims, or in some cases was collected by using hashtags or keywords in searching such as Islam, Muslim, Quran, Muhammed, hadith and Sharia, and their synonyms. We used the term “Disinformation" because we deal with facts and information related to the Islamic religion. We would like to point out to the non-Muslim reader that what is meant by the “Qur’an” is the holy book of Islam and “Hadith” is the sayings and actions of the prophet Muhammed. Sharia is a set of religious laws that are part of the Islamic tradition. It is derived from the religious teachings of Islam and is based on the holy books of Islam, especially the Qur'an, and Hadith [57].

4.1.2. Dataset Labeling

The ground truth data for correct information about the Islamic religion was collected from the main sources of Islamic rulings and legislation. This data was verified by a team of Islamic scholars. Disinformation targeting the Islamic religion was collected from Facebook pages and groups and verified using reliable Islamic websites. The data relating to the Islamic religion has been manually labeled under the supervision of Islamic domain experts. The Islamic scholars' team consists of two Imams (The leader of worshipers in a mosque is usually called the imam) and a professor, who hold Ph.D. degrees in Islamic creed, interpretation of the Quran, and Hadith. As well as reliable Islamic web pages for fatwas (Islamic judgment) were used such as Islamweb.net (https://www.islamweb.net/en/ , accessed on 16 July 2021) and Islam Question & Answer (https://islamqa.info/en, accessed on 20 July 2021) webpages. The Islamic scholars investigated the dataset in terms of the correlation of ground-truth data with the scope of research, and also the reliability of fatwa webpages used to verify the disinformation. As was mentioned in the introduction the difference between Misinformation and Disinformation is based on the intention or purpose of posting. Since all the misleading information was extracted from pages or groups intended to offend or distort the Islamic religion, most of these posts were posted by admins and were not shared by individuals. all these allegations were considered Disinformation. It should be noted that any information or article that is modified or misleading information added to it has been considered disinformation. Since we are dealing with a critical issue that does not accept fine-grained classifications, the classification adopted in this research is binary, meaning either real or false. This is even if it contains facts or cites some Quranic verses or hadiths.

4.1.3. Dataset Description

The CIDII dataset is a binary classification, consisting of the two classes of correct information and disinformation This dataset is available in our account on Google drive in CSV format (https://drive.google.com/drive/folders/14HBFsDAe8hbNn6S6qTJ84o-9_IETDOq0?usp=sharing, accessed on 12 January 2023). Table 3 illustrates the structure and content description of the dataset.

Table 3: The description of the CIDII dataset

Column No.	Dataset Columns Names	Description
1	ID	Each article has a unique ID.
2	Article	The article contains text that is either facts related to Islamic issues if the information is correct, or posts targeting the Islamic religion if the information is false. Most posts contain only the body without a title.
3	Propagation Source	The source refers to the source of the article content, as it contains a Facebook link in the event that the post is disinformation, or it contains a link to Islamic websites in the event that the article refers to correct information (an explanation of a verse, a hadith, or an article related to the Islamic religion).
5	Article Type	This column contains the type of article published. Is it a post if the article is disinformation, or is it an Islamic article, a Quranic interpretation, or a hadith, in the event that the information is correct?
6	Class Type	This column shows whether the article belongs to the category of correct information or disinformation.

4.1.4. Dataset Visualization

We implemented exploratory data analysis on the gathered dataset before performing any pre-processing step to obtain some general perception. Figures 4 and 5 illustrate the word cloud of the top frequent words in the ground truth data and disinformation posts. We observe from these collected data that the vocabulary found belongs to a domain-specific language, and differs from the rest of the vocabulary used in other domains. Figures 6 and 7 show the top 20 words used in the correct information and disinformation classes after removing stop words. We note through these two Figures 6 and 7, that despite the presence of close-in numbers in the repeated words in the two classes of correct and fake information, the majority of the disinformation revolves around explicit names (Muhammad, Islam, Allah, Muslims, Qur’an). As for the correct information, we see the most repeated word is the word (Allah), but concerning the Prophet Muhammad, he is not mentioned by his explicit name (Muhammad), but rather by his attributes as a (prophet) or (messenger). Figures 8, and 9, represent the lengths of articles of two types of collected data for the CIDII dataset. These articles show that articles of correct information are slightly longer than disinformation articles. Figures 10 and 11 show the number of words in these articles from the CIDII dataset. As can be seen from these two Figures 10 and 11, correct information articles also contain more words than disinformation. Figures 12 and 13, present the top 20 Part-Of-Speech tagging in the correct information and disinformation classes of the CIDII dataset. As shown in Figures 12 and 13, correct information is composed of (singular or plural nouns, prepositions, numbers, and determinants) respectively. In contrast, disinformation often contains singular proper nouns and then singular or plural nouns. In Figures 14 and 15, we notice that the polarity of sentiment for the two classes of the dataset is close to each other. Most of them fall into the neutral sentiment category. Figures 16, 17, 18, 19, and 20 indicate that disinformation articles related to the Islamic domain are characterized by containing words in capital letters for amplification or including exclamation marks and question marks to astonish the reader. Additionally, by reviewing this dataset, the disinformation articles are poorly worded and contain slang, misspellings, profanity, or unusual repetition of characters to attract attention. The tools and code that were used to analyze the data and extract these results and Figures are available in our Google Colab platform account (https://colab.research.google.com/drive/1CqrYVH3tb_1r9JWt5e_NTMssczP5LPti?usp=sharing, accessed on 13 January 2023).

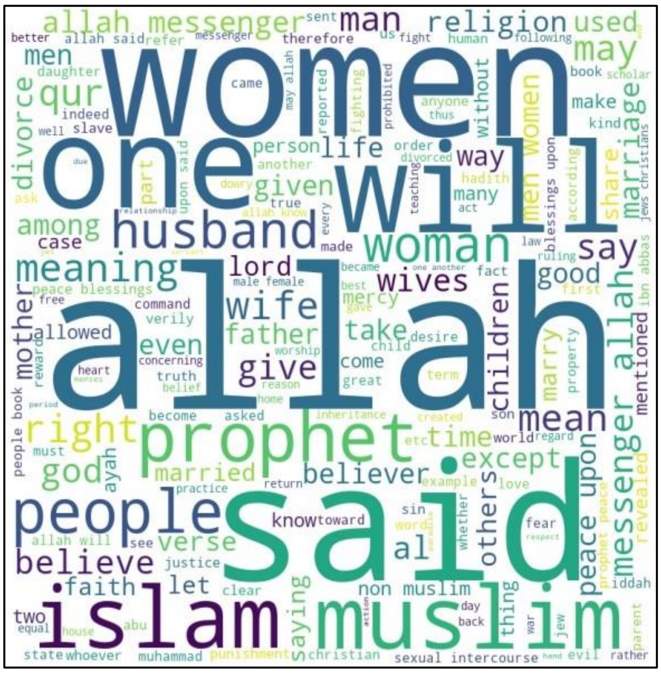


Fig. 4: Word cloud visualization of correct information

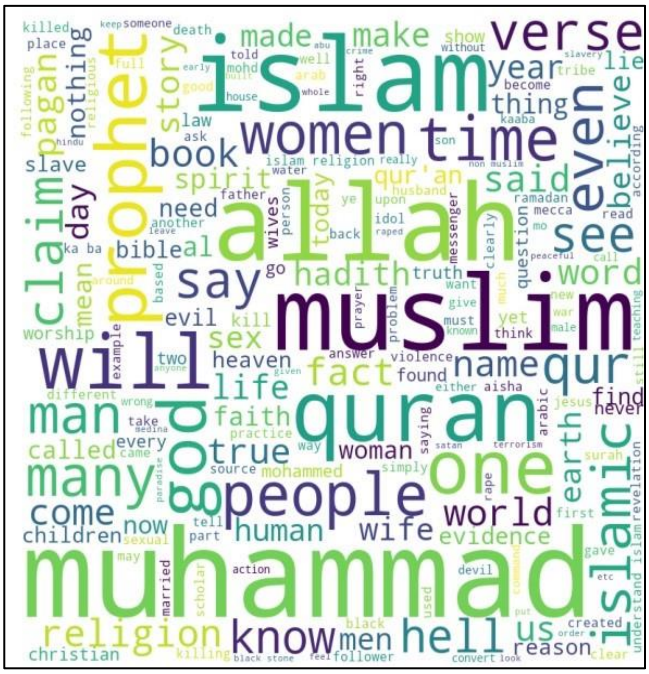


Fig. 5: Word cloud visualization of disinformation

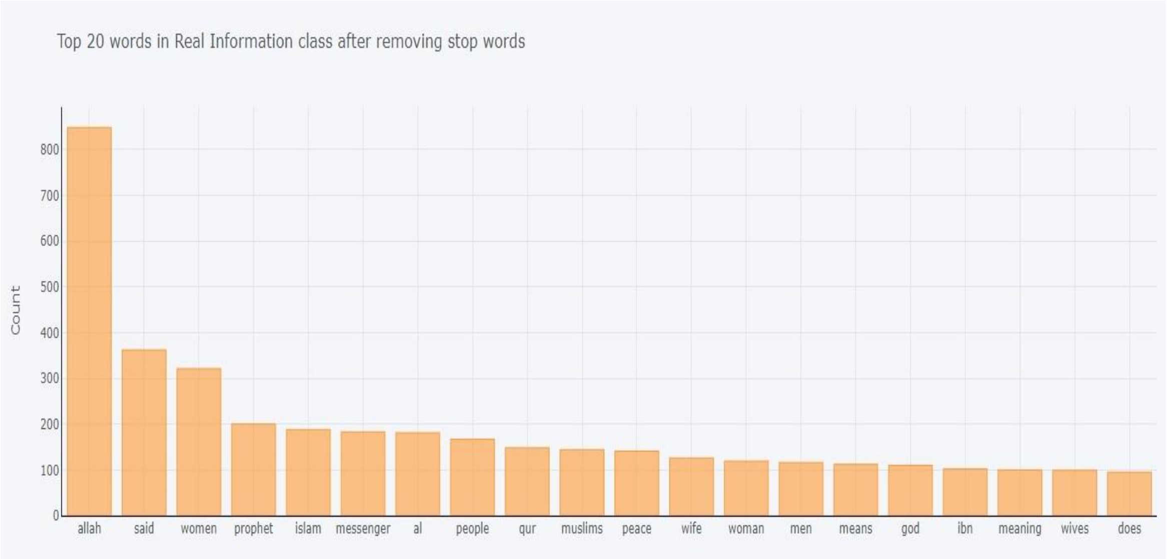


Fig. 6: Top 20 words in correct information class after removing stop words

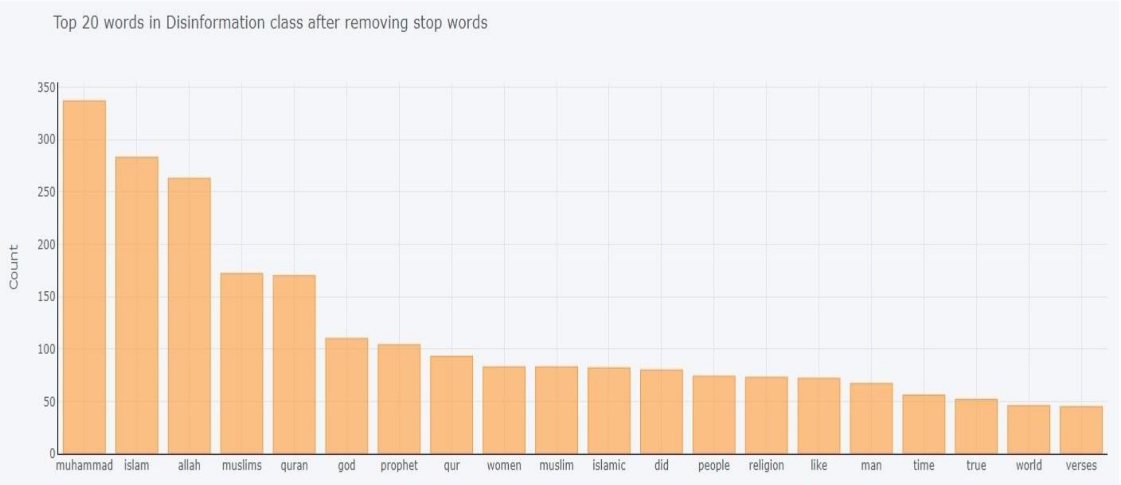


Fig. 7: Top 20 words in disinformation class after removing stop words

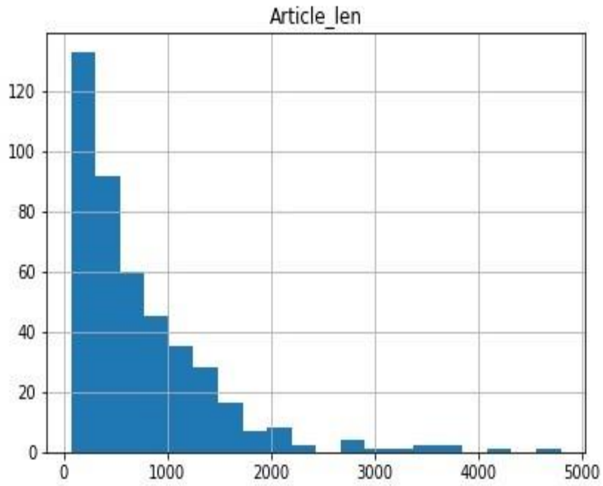


Fig. 8:The article lengths for the correct information class

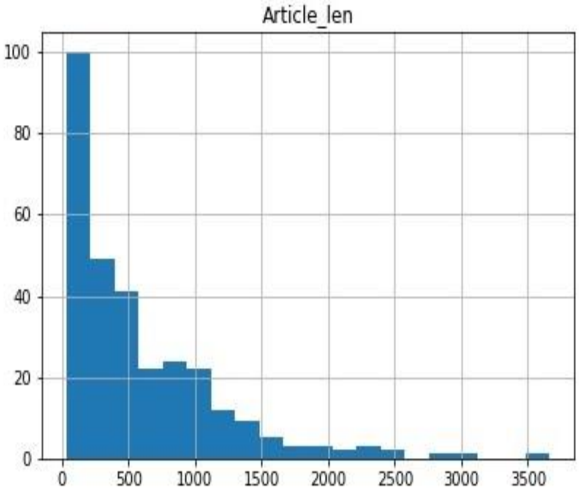


Fig. 9: The article lengths for the disinformation class

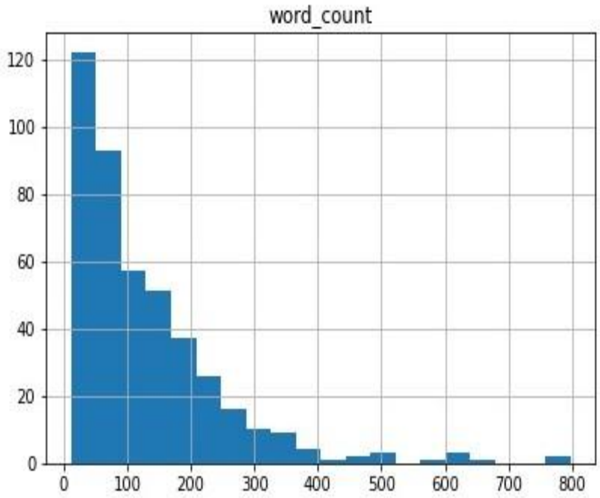


Fig. 10: The number of words in correct information articles.

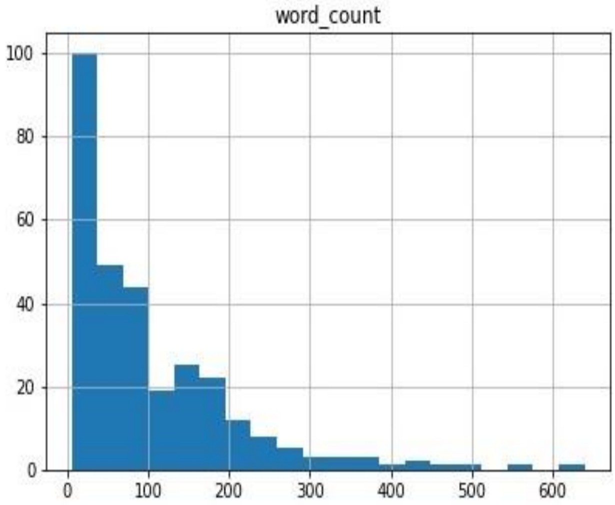


Fig. 11: The number of words in disinformation articles.

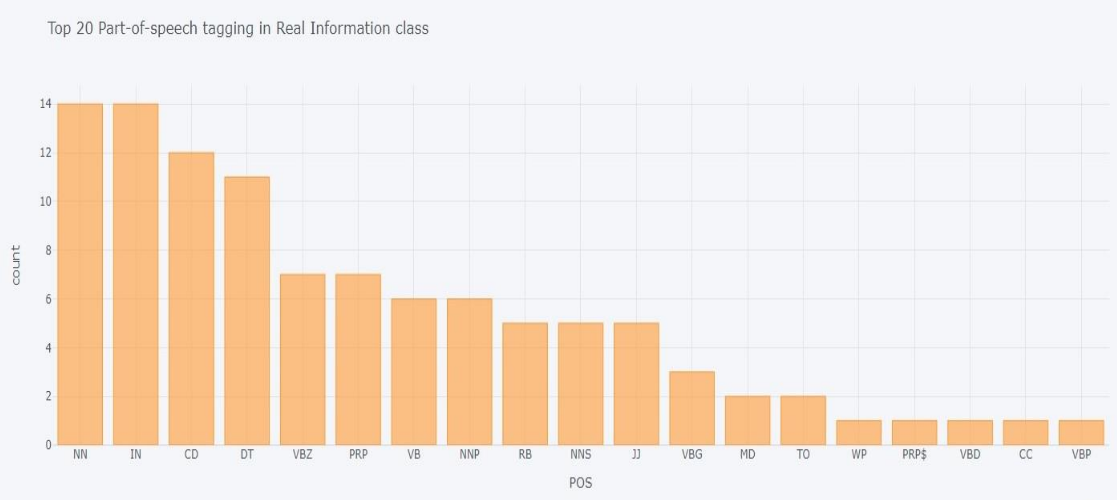


Fig. 12: The top 20 Part-Of-Speech tags in the correct information class

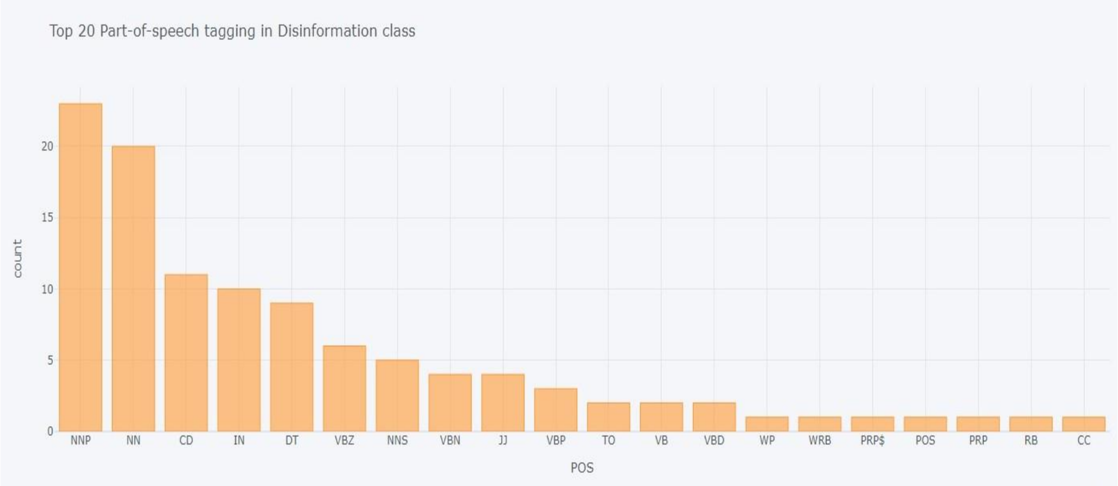


Fig. 13: The top 20 Part-Of-Speech tags in the disinformation class

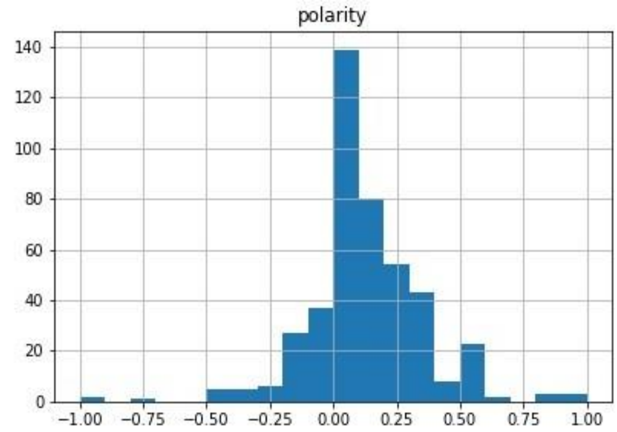


Fig. 14: The sentiment polarity of the correct information class

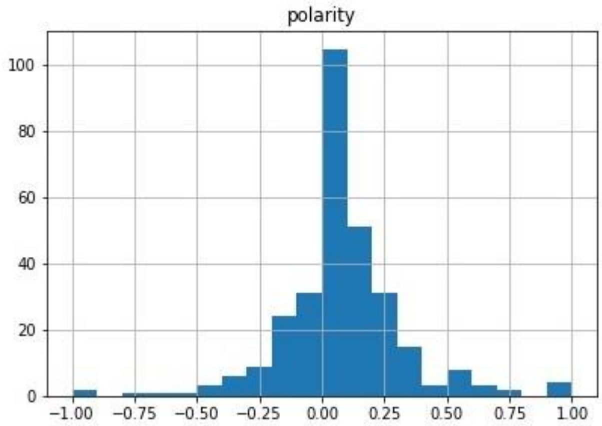


Fig. 15: The sentiment polarity of the disinformation class

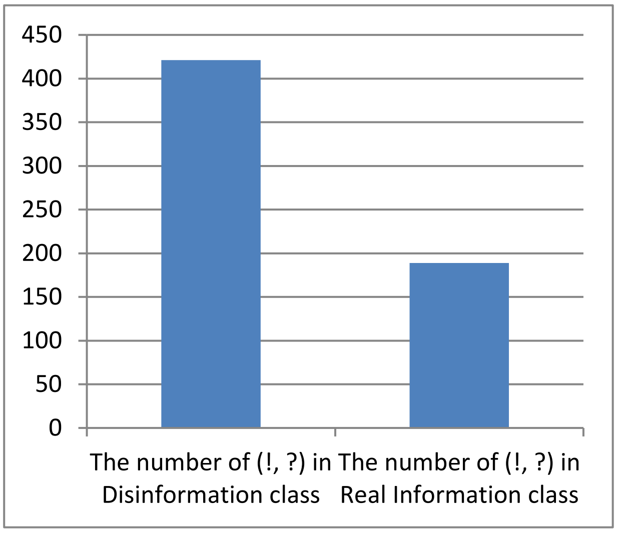


Fig. 16: The number of exclamation and question marks in the classes of correct information and disinformation

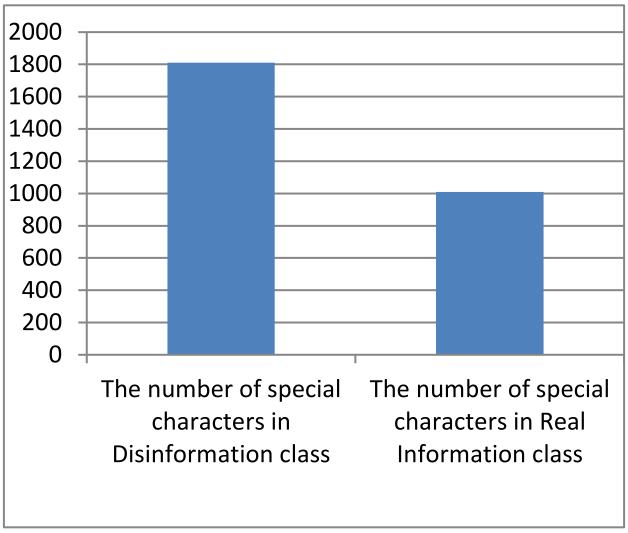


Fig. 17: The number of special characters in the classes of correct information and disinformation

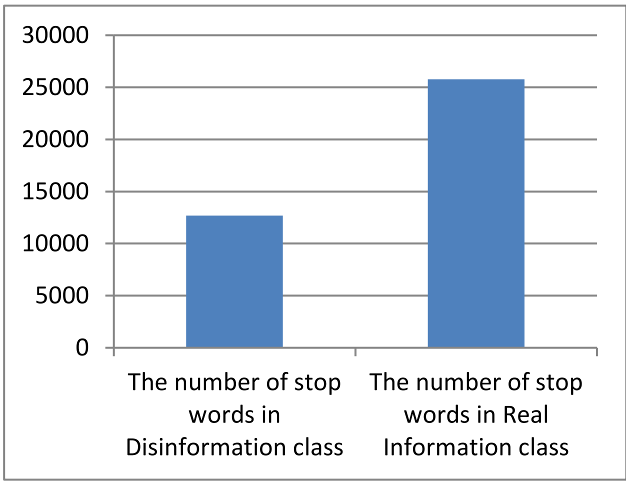


Fig. 18: The number of stop words in the classes of correct information and disinformation

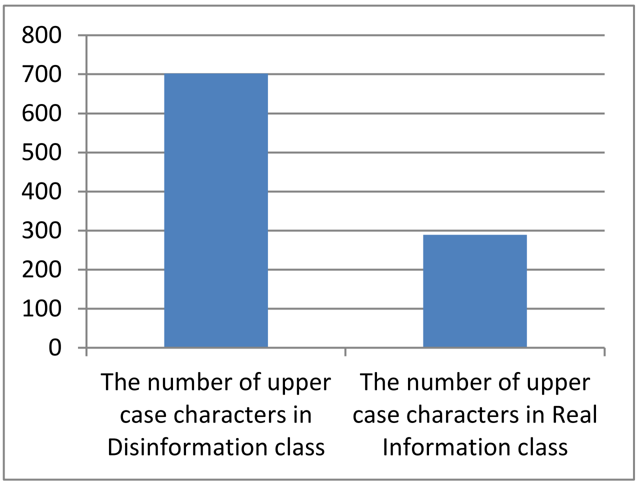


Fig. 19: The number of upper-case characters in the classes of correct information and disinformation

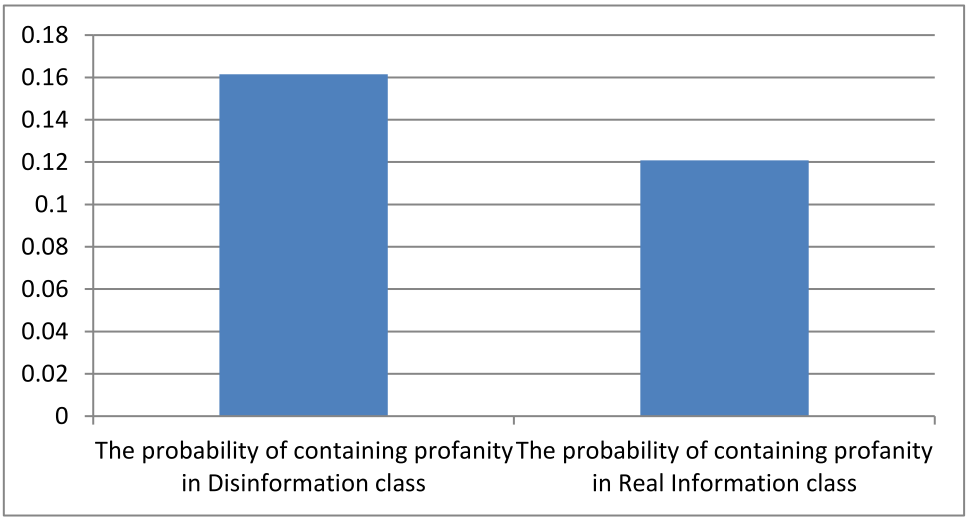


Fig. 20: The number of profanity words in the classes of the CIDII dataset

Table 4: Summary of some textual attributes in the CIDII dataset

Attribute	Correct Information class	Disinformation class
Number of stop words	25788	12672
Number of upper-case characters	289	702
Number of special characters	1009	1811
Number of (!, ?)	189	421
The probability of containing profanity	0.12%	0.16%

4.2 The Proposed Detection Model

The proposed disinformation detection model will determine whether the post is Correct or Disinformation depending on the content features. This model relies on the Bi-LSTM classifier to classify Islamic articles based on labeled instances. The text inputs are embedded using pre-trained GloVe word embedding which will be retrained on Islamic documents as illustrated in Figure 21. The proposed detection model section is divided into the following subsections: 4.2.1. Text Pre-Processing Phase, 4.2.2. Feature Extraction Phase based on Word Embedding, and 4.2.3. Disinformation Detection Phase.

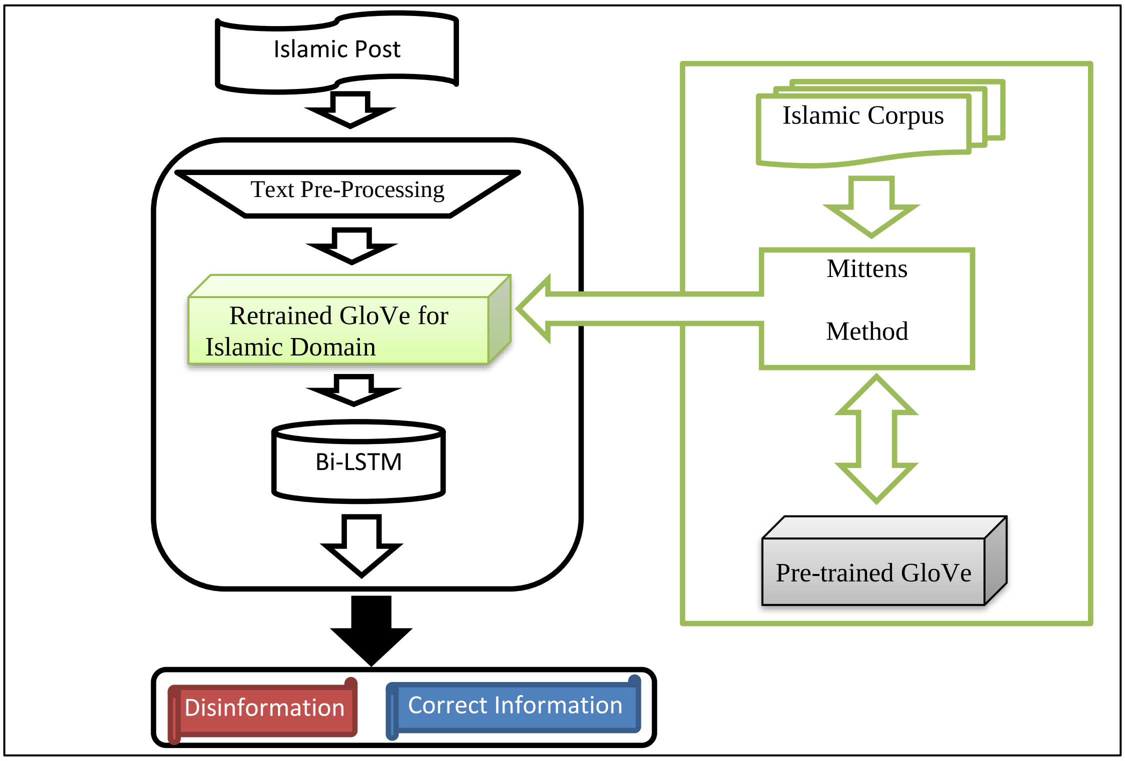


Fig. 21: The general design of the proposed disinformation detection model related to the Islamic domain

4.2.1. Text Pre-Processing Phase

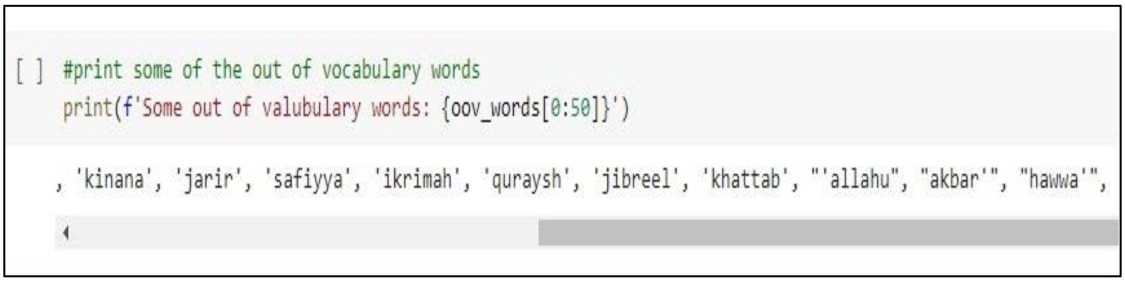
Input preprocessing is considered the first part of our model. Since the proposed model uses a deep learning technique, the style of writing is one of the features that are taken into account in detecting disinformation [42]. Consequently, the pre-processing phase of this model will be limited to the following steps as shown in Algorithm 1.

```
Algorithm 1: Text preprocessing
Input:      CSV files
Output:     Preprocessed dataset
Procedure:  Preprocess Post(text):
    Read input data (text)
    For each input (text)
    Begin
        text = remove links(text)
        text = remove accented characters(text)
        text = remove duplicate characters(text)
        text = expand contractions(text)
        text = remove special characters except exclamation question(text)
        text = remove stopwords(text)
        text = remove white spaces(text)
        tokens = tokenized text(text)
        padded tokens = pad tokens(tokens)
    End
    Save the processed data (padded tokens)
```

After performing the above-mentioned preprocessing steps, let $x_{1,}\ x_{2,}x_{3,}\dots,\ x_v$ represent all the unique words in the dictionary. D= $d_{1,}\ d_{2,}d_{3,}\dots,\ d_m$ then $i_{1,}\ i_{2,}i_{3,}\dots,\ i_v$ be compatible with the number of distinct indices. The symbols 1 and V indicate the first and last indexes in the dictionary of vocabulary, respectively, whereas indices denote natural numbers. The data input is carried in the input phase as a series of identical-length unique indices.

4.2.2. Feature Extraction Phase based on Word Embedding

The process of feature extraction, which converts each text into a numerical representation in the form of a vector [58], is the initial stage of training a deep-learning model. One of the common and effective techniques used in deep learning models for feature extraction is word embedding. Each word is converted into a vector via word embedding models, which identify relationships and similarities between words in the corpus of text. In other words, word embeddings pack more data into fewer dimensions. It should be noted that word embeddings map the statistical structure of the language used in the corpus rather than understanding the text as a human would. The goal of word embedding is to map semantic meaning into the embedding space [59]. This would map words with close semantic similarities to the embedding space, such as cities, colors, and numbers [34]. The pre-trained GloVe model is a type of word embedding that handles high-dimensional news articles. Stanford created the GloVe model [60], which is superior to the Word2Vec model since it is trained using global iteration numbers as opposed to discrete local context windows in Word2Vec [61]. The pre-trained GloVe word embedding model could be used to determine how two words are related based on how far apart they are in a vector space [62]. We will use the cased pre-trained GloVe model (840B tokens, 2.2M vocabulary, 300D vectors) (https://nlp.stanford.edu/projects/glove/, accessed on 23 September 2021), to handle both cases of the word in the case of writing in uppercase or lowercase, and this is because capitalization of words is common in disinformation. Another aspect element that affects the process of disinformation detection is vocabulary size [63]. Since we are dealing with a specific domain, where this domain contains Islamic vocabulary that is not commonly found in other domains, most pre-trained word embedding models do not have such terms. Moreover, there are entity names related to newly introduced events that may not be included in these models. Accordingly, the problem of Out-Of-Vocabulary (OOV) words will arise during the training process. These OOV words do not have any value in the GloVe model, that is, these words exist in the dataset used for the Islamic issues domain and do not exist in the pre-trained GloVe model. Thus, these words will then be represented by zeros. This pre-trained GloVe model will be retrained to Islamic documents to solve this problem. Figure 23 is a screenshot of OOV words that were encountered in the model during the training phase before retraining the pre-trained GloVe. With the use of a specific word embedding model, we can achieve a better representation of the linguistic features of our dataset [63]. This section contains two subsections. The first section (i) describes the corpus used to retrain the pre-trained GloVe model, and the second section (ii) shows how the GloVe model will be retrained using the Mitten method.



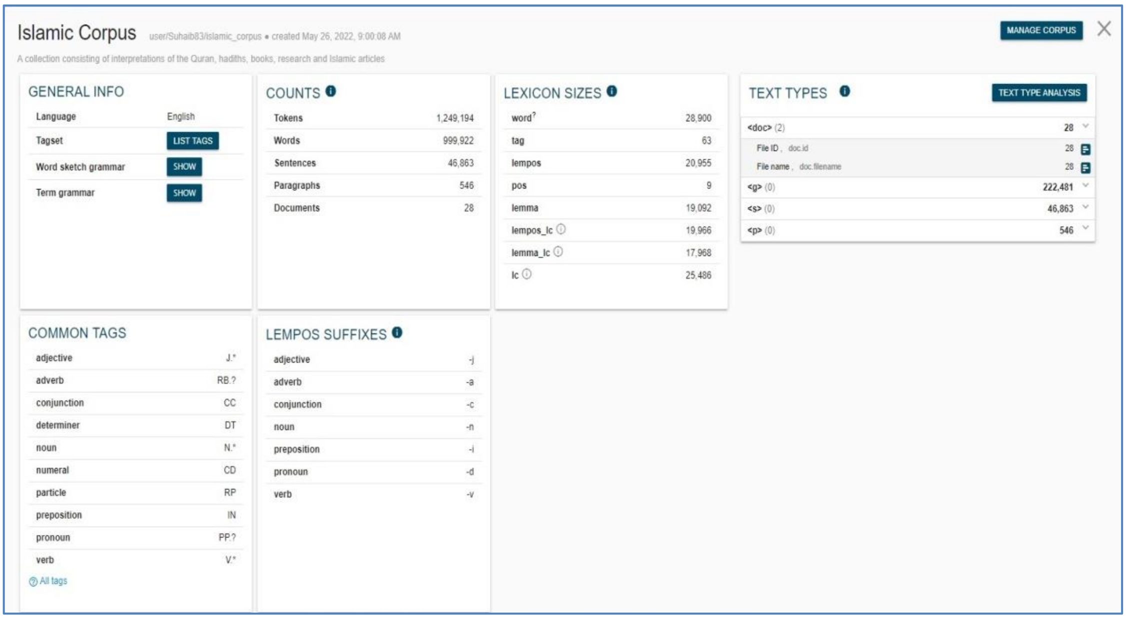
```
[ ] #print some of the out of vocabulary words
    print(f'Some out of valubulary words: {oov_words[0:50]}')

  , 'kinana', 'jarir', 'safiyya', 'ikrimah', 'quraysh', 'jibreel', 'khattab', "'allahu", "akbar'", "hawwa'",
```

Fig. 22: A screenshot of generated Out-Of-Vocabulary words

I) Islamic Corpus

An Islamic corpus was created using the sketch engine tool (https://www.sketchengine.eu/, accessed on 4 October 2021). This corpus consists of 999,922 words in general and 28,900 unique words extracted from 18 Islamic articles from the Islam Online webpage. In addition, 10 Islamic documents include the interpretation of the Qur'an, hadith in English, and some other Islamic documents. This corpus can be accessed through our Google drive account (https://drive.google.com/drive/folders/14UQmz2-sJEz23SFMq2-qXi-cS-hEASkc?usp=sharing, accessed on 5 January 2023). Figure 24 shows information from the Islamic corpus.



Islamic Corpus user/Suhaib83/islamic_corpus • created May 26, 2022, 9:00:08 AM

A collection consisting of interpretations of the Quran, hadiths, books, research and Islamic articles

GENERAL INFO

Language	English
Tagset	LIST TAGS
Word sketch grammar	SHOW
Term grammar	SHOW

COUNTS

Tokens	1,249,194
Words	999,922
Sentences	46,863
Paragraphs	546
Documents	28

LEXICON SIZES

word?	28,900
tag	63
lempos	20,955
pos	9
lemma	19,092
lempos_lc	19,966
lemma_lc	17,968
lc	25,486

TEXT TYPES

<doc> (2)	28
File ID , doc.id	28
File name , doc.filename	28
<g> (0)	222,481
<s> (0)	46,863
<p> (0)	546

COMMON TAGS

adjective	J.*
adverb	RB.?
conjunction	CC
determiner	DT
noun	N.*
numeral	CD
particle	RP
preposition	IN
pronoun	PP.?
verb	V.*

LEMPOS SUFFIXES

adjective	-j
adverb	-a
conjunction	-c
noun	-n
preposition	-i
pronoun	-d
verb	-v

Fig. 23: A screenshot of Information from the Islamic corpus based on sketch engine software [64]

II) Retraining GloVe based on Mittens

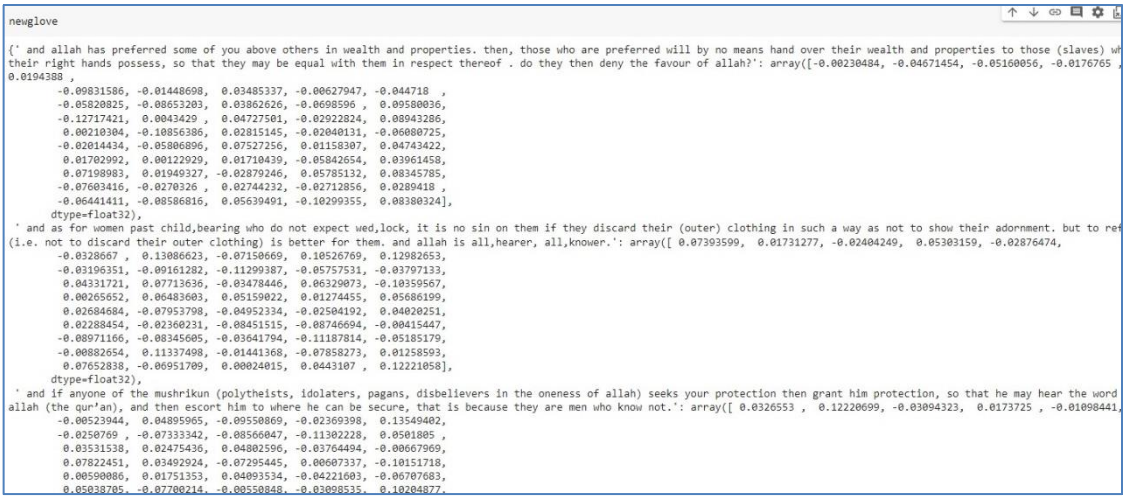
An Islamic Corpus was used to retrain the pre-trained GloVe model using the Mittens approach of Dingwall and Potts [65]. Mittens is a simplified extension of the pre-trained GloVe model that updates standard pre-trained GloVe representations based on data from a specific domain. To retrain the GloVe model based on Mittens, firstly, the pre-trained GloVe model had to be loaded as a dictionary for Mittens. Before creating the co-occurrence matrix of the words, we shall pre-process the Islamic corpus, using stop word removal, normalization, and padding techniques. The OOV words are used to create the co-occurrence matrix. OOV refers to terms missing from the pre-trained GloVe. To build a co-occurrence matrix, a word-word co-occurrence is what we require, not the regular term-document matrix. The word-document matrix is created from the document using Sklearn's CountVectorizer (https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html, accessed on 18 November 2022). Finally, We train Mittens (Mittens is available as a Python library)(https://github.com/roamanalytics/mittens, accessed on 7 December 2022) because we have our embeddings (stored as original embeddings), a co-occurrence matrix, and related vocabulary. When new embeddings are trained, they should be compatible with the current embeddings. This is because they will be oriented in a way that makes employing a combination of the two embeddings relevant. The Algorithm 2 summarize the process of retraining the GloVe model based on Mittens.

Algorithm 2: Retrained GloVe based on Mittens

```
Input:      Islamic Corpus
Output:     Retrained GloVe Model
Procedure:
1   Load_pretrained_model():
            Load the pre-trained GloVe model as a dictionary for Mittens
            Return the loaded model.
2   Preprocess_corpus(corpus):
            Remove stop words from the corpus.
            Tokenize the corpus.
            Pad the corpus.
            Return the preprocessed corpus.
3   Build_co_occurrence_matrix(corpus):
            Create a word-word co-occurrence matrix from the corpus.
            Return the co-occurrence matrix.
4   Train_Mittens(GloVe_model, co_occurrence_matrix):
            Train the Mittens model using the pre-trained GloVe model and co-occurrence matrix.
            Return the retrained GloVe model.
5   Main():
            GloVe_model = Load_pretrained_model()
            Preprocessed_corpus = Preprocess_corpus(Islamic_Corpus)
            Co_occurrence_matrix = Build_co_occurrence_matrix (Preprocessed_corpus)
            Retrained_model = Train_Mittens(GloVe_model, Co_occurrence_matrix)
            Return Retrained_model
```

- **Load the pre-trained GloVe model:** The pre-trained GloVe model had to be loaded as a dictionary for Mittens.
- **Data pre-processing:** Before creating the co-occurrence matrix of the words, we shall pre-process the Islamic corpus, using stop word removal, normalization, and padding techniques. The OOV words are used to create the co-occurrence matrix. OOV refers to terms missing from the pre-trained GloVe.
- **Build a co-occurrence matrix:** Word-word co-occurrence is what we require, not the regular term-document matrix. The word-document matrix is created from the document using Sklearn's CountVectorizer (https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html, accessed on 18 November 2022).
- **Retraining the GloVe model:** We train Mittens (Mittens is available as a Python library) (https://github.com/roamanalytics/mittens, accessed on 7 December 2022) because we have our embeddings (stored as original embeddings), a co-occurrence matrix, and related vocabulary. When new embeddings are trained, they should be compatible with the current embeddings. This is because they will be oriented in a way that makes employing a combination of the two embeddings relevant.

The code for the steps to retrain the GloVe model is available in our Google Colab account (https://colab.research.google.com/drive/10ntcC8V7GMf2vFk0_H2O3mdfKYAbCZqP?usp=sharing , accessed on 12 January 2023). As we can see in Figure 25, a vector with a size of 10 x 5 (50 D) represents the numeric vectors of OOV words embedded. The newly retrained GloVe model will be used in the proposed model through the embedding layer.



```
newglove

{' and allah has preferred some of you above others in wealth and properties. then, those who are preferred will by no means hand over their wealth and properties to those (slaves) wh
their right hands possess, so that they may be equal with them in respect thereof . do they then deny the favour of allah?': array([-0.00230484, -0.04671454, -0.05160056, -0.0176765 ,
0.0194388 ,
        -0.09831586, -0.01448698,  0.03485337, -0.00627947, -0.044718  ,
        -0.05820825, -0.08653203,  0.03862626, -0.0698596 ,  0.09580036,
        -0.12717421,  0.0043429 ,  0.04727501, -0.02922824,  0.08943286,
         0.00210304, -0.10856386,  0.02815145, -0.02040131, -0.06080725,
        -0.02014434, -0.05806896,  0.07527256,  0.01158307,  0.04743422,
         0.01702992,  0.00122929,  0.01710439, -0.05842654,  0.03961458,
         0.07198983,  0.01949327, -0.02879246,  0.05785132,  0.08345785,
        -0.07603416, -0.0270326 ,  0.02744232, -0.02712856,  0.0289418 ,
        -0.06441411, -0.08586816,  0.05639491, -0.10299355,  0.08380324],
       dtype=float32),
 ' and as for women past child,bearing who do not expect wed,lock, it is no sin on them if they discard their (outer) clothing in such a way as not to show their adornment. but to ref
(i.e. not to discard their outer clothing) is better for them. and allah is all,hearer, all,knower.': array([ 0.07393599,  0.01731277, -0.02404249,  0.05303159, -0.02876474,
        -0.0328667 ,  0.13086623, -0.07150669,  0.10526769,  0.12982653,
        -0.03196351, -0.09161282, -0.11299387, -0.05757531, -0.03797133,
         0.04331721,  0.07713636, -0.03478446,  0.06329073, -0.10359567,
         0.00265652,  0.06483603,  0.05159022,  0.01274455,  0.05686199,
         0.02684684, -0.07953798, -0.04952334, -0.02504192,  0.04020251,
         0.02288454, -0.02360231, -0.08451515, -0.08746694, -0.00415447,
        -0.08971166, -0.08345605, -0.03641794, -0.11187814, -0.05185179,
        -0.00882654,  0.11337498, -0.01441368, -0.07858273,  0.01258593,
         0.07652838, -0.06951709,  0.00024015,  0.0443107 ,  0.12221058],
       dtype=float32),
 ' and if anyone of the mushrikun (polytheists, idolaters, pagans, disbelievers in the oneness of allah) seeks your protection then grant him protection, so that he may hear the word
allah (the qur'an), and then escort him to where he can be secure, that is because they are men who know not.': array([ 0.0326553 ,  0.12220699, -0.03094323,  0.0173725 , -0.01098441,
        -0.00523944,  0.04895965, -0.09550869, -0.02369398,  0.13549402,
        -0.0250769 , -0.07333342, -0.08566047, -0.11302228,  0.0501805 ,
         0.03531538,  0.02475436,  0.04802596, -0.03764494, -0.00667969,
         0.07822451,  0.03492924, -0.07295445,  0.00607337, -0.10151718,
         0.00590086,  0.01751353,  0.04093534, -0.04221603, -0.06707683,
         0.05038705, -0.07700214, -0.00550848, -0.03098535,  0.10204877,
```

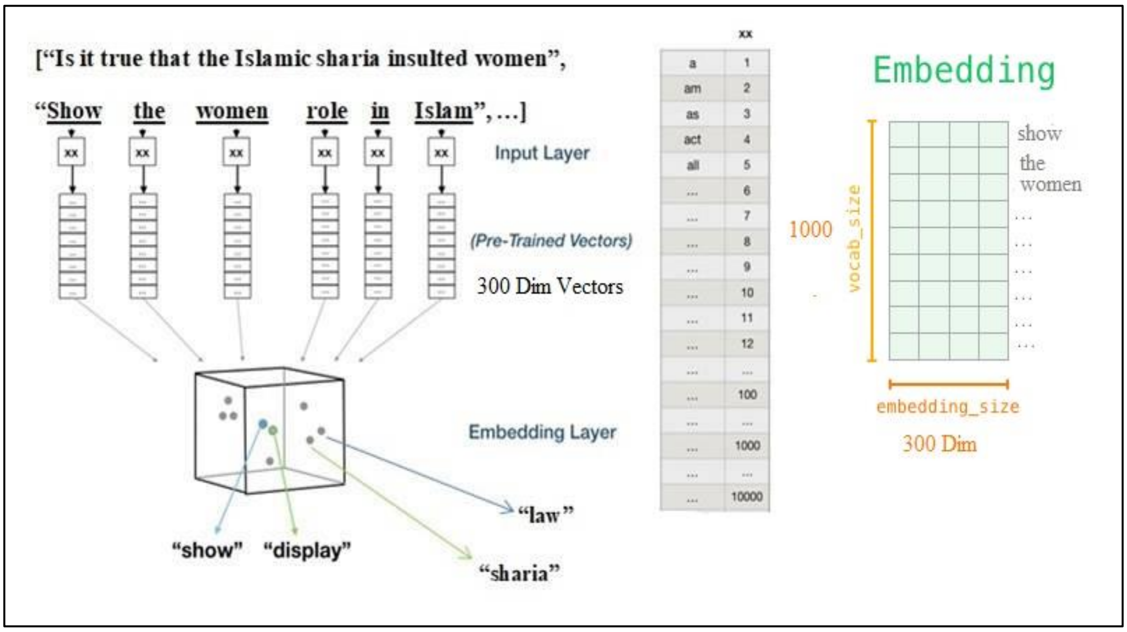
Fig. 24: A screenshot of new word embeddings for Islamic terms in the retrained GloVe model

4.2.3. Disinformation Detection Phase

This phase is responsible for detecting disinformation about Islam based on training the model on labeled examples of articles. The detection of disinformation in this research is a binary classification. Each article a_i is classified as a label *C*, where a_i belongs to (A= $a_{1,}$ $a_{2,}a_{3,}\ldots$, a_m) where *m* is the number of articles in the dataset, and *C* represents one of two predefined classes (*C*= 2). This model consists of 3 main layers: the embedding layer, the Bi-LSTM layer, and the dense layer. This section contains two subsections. Section (i) Embedding Layer, and Section (ii) Bidirectional LSTM Classifier.

I) The Embedding Layer

Every index that corresponds to a distinct word in the dataset is converted into a real-valued feature vector in the embedding layer. These real-valued vectors are piled up to create an embedding matrix. The embedding matrix is designed with the idea that each row represents a distinct index, which in turn correlates to a distinct word in the dictionary of vocabulary. The embedding matrix's dimension is $v \times d$, where *v* represents the vocabulary size of the dataset and *d* represents the size of the dense vector. In our study, we employed a pre-trained GloVe word embedding model that was based on a 300-dimensional vector. Since we retrained the pre-trained GloVe model based on Islamic documents using the Mittens approach, therefore, the embed layer of the model loads initial values (vectors) for the words from the retrained GloVe embedding of the Islamic domain instead of loading random representations [50] as shown in Figure 26, and will be fed to the next layer of the classifier.



	xx
a	1
am	2
as	3
act	4
all	5
...	6
...	7
...	8
...	9
...	10
...	11
...	12
...	...
...	100
...	...
...	1000
...	...
...	10000

Fig. 25: Vector representation of words based on retrained GloVe embedding

II) Bidirectional LSTM Classifier

Deep Learning classifiers learn to make classifications based on prior observations rather than depending on human-created rules. DL techniques can understand the different associations between textual fragments and that a specific output (i.e., labels) is predicted for a specific input by using pre-labeled examples as training data. For modeling the pertinent features of disinformation, the bidirectional training method is preferred because it can improve classification performance while capturing semantic and long-distance connections in sentences [62]. Bi-LSTM is an advancement over the regular LSTM that can increase the model's effectiveness through training two isolated LSTMs [20]. For the prediction and classification of large text sequences, bidirectional processing is an excellent strategy [50], as the collected dataset includes long articles. Bi-LSTM consists of a forward LSTM layer and a back LSTM layer. The front layer captures the historical information of the sequence. This information is processed by the forward LSTM from left to right, and its hidden state can be represented as $\overrightarrow{ht}$ = LSTM (x_t , $\overrightarrow{ht}$-1). The back layer captures the future information of the sequence. The backward LSTM will process the information from right to left and its hidden state is denoted as $\overleftarrow{ht}$ = LSTM (x_t , $\overleftarrow{ht}$+1). These front and back layers are connected to the output layer. The forward and backward states can be combined as h_t = [$\overrightarrow{ht}$, $\overleftarrow{ht}$] to represent the output of Bi-LSTM. Finally, the dense layer, which is a fully connected layer based on the Sigmoid activation function will classify the articles into two classes: correct information, and disinformation, as illustrated in Figure 27. Due to the binary classification involved, the Sigmoid activation function is the appropriate choice. It can be represented as follows:

$$F(x) = \frac{1}{1+e^{-x}} \tag{1}$$

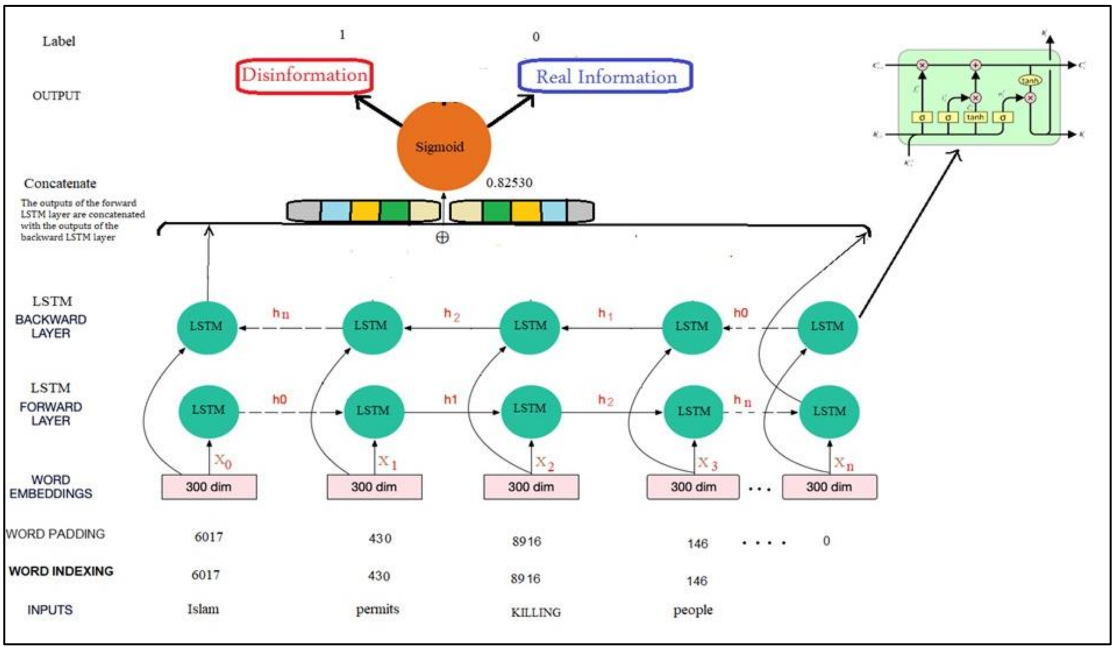


Fig. 26: The structure of the Bi-LSTM Model.

The Bi-LSTM classifier has been trained and validated on the RIDI dataset. One of the key components of any deep learning solution is the selection of appropriate hyperparameters. The balanced or imbalanced dataset determines how to choose the best values for hyperparameters. There are two basic methods for selecting optimal values: automatic selection and manual selection. Both approaches are equally legitimate, but manual selection requires a thorough understanding of the model. Our model's hyperparameters were chosen based on the most successful outcomes as shown in the following Table 4 after we examined and analyzed the results of multiple classifiers with various learning paradigms (different optimal hyperparameters and architectures) [62]. We used the Sigmoid function as an activation function and binary cross entropy as a loss function because we deal with binary classification.

Table 5: The architecture of the disinformation detection model about the Islamic domain

No.	**Structure or Hyperparameter Name**	**Type or Value**
1	Layers	Embedding layer Bidirectional LSTM layer Dense layer
2	Word embedding dimension	300
3	No. of hidden states	256
4	Dropout	0.2
5	Recurrent dropout	0.2
6	Activation function	Sigmoid
7	Loss function	Binary_Crossentropy
8	Optimizer	Adam
9	Learning rate	0.1
10	Batch size	256
11	No. of epochs	5

4.3 Performance Evaluation

The field of fake news detection is one of the new fields, which contains few standard datasets [66], [67]. Most of this dataset is devoted to political, electoral, or financial data. There is no standard dataset that investigates the topic of detecting fake news or disinformation in the Islamic domain. Therefore, to evaluate the performance of the proposed model, we will compare it with baseline models that are highly efficient at dealing with sequential text. These models include LSTM and GRU models [68], [48]. In this research, we used 80% of the dataset for model training and 10% for the validation, and 10% for the model test. Since we have an imbalanced dataset, therefore, to evaluate the performance of the model, the Area Under Curve (AUC) performance measure was applied. It provides more accurate performance measures for imbalanced datasets when compared with the F1-score [4]. We will use the AUC measure as well as the F1-score metric. AUC is a measure of a classifier's ability to distinguish between classes, where the higher the value of the AUC, the better the performance of the model. AUC measurement is used to compare learning algorithms and build optimum learning models. Its value reflects how well a classifier performs overall in terms of ranking. Equation 2 represents the formula for AUC.

$$\mathbf{AUC} = \frac{\mathbf{1-FPR+TPR}}{\mathbf{2}} \qquad 2$$

Where FPR indicates False Positive Rate, which is the proportion of negatively classified instances to all instances of negative instances. True Positive Rate, or TPR, is an acronym that stands for the percentage of positive examples that are accurately classified [53]. Regarding the F1-score, the following Equation 3 shows the formula for the F1-score:

$$\mathbf{F1-Score} = \frac{\mathbf{2 \times Precision \times Recall}}{\mathbf{Precision+\ Recall}} \qquad 3$$

F1-Score is the result of the harmonic mean of Precision and recall, where Precision is the ratio of correctly predicted positive instances to all other positive instances in a class of predictions. As with recall, it is used to calculate the percentage of correctly categorized positive examples.

5.0 EXPERIMENTAL RESULTS AND DISCUSSION

We conducted the experiments using the Python Programming Language on the Google Colab platform (https://colab.research.google.com/, accessed on 12 January 2023), and the full code of all the experiments that were performed is available on our Google Colab account (https://colab.research.google.com/drive/1JFPsR0GIo-F0nNGG8Q1DWskNMKiPmyvy?usp=sharing, accessed on 12 January 2023). It is known that a feature-rich dataset significantly affects the performance of the model since most of the available standard datasets contain examples that are either general news in various domains or are specific to a particular domain, such as politics, economics, elections, or COVID-19. Therefore, to show the importance of providing a dataset related to the Islamic domain to train a model capable of detecting disinformation related to the Islamic domain, we trained our model on two different datasets: The Fake or Real News dataset (https://www.kaggle.com/rchitic17/real-or-fake, accessed on 11 May 2021), which contains political and social news; and the CIDII dataset, which was collected. We tested these two trained models on the CIDII dataset. As shown in Table 5, the model trained on the general dataset failed to detect disinformation related to the Islamic religion. It provided poor accuracy based on the AUC measure, which is 51.49%. This result is plausible because the two datasets used in the test are different in terms of features, domain language, and writing style. This is depicted in Figures 28 and 29.

Table 6: Comparison of the results of a model trained on two different datasets

Model	Training Dataset Name	Dataset Size	Training AUC Measure	Testing Dataset Name	Testing AUC Measure
Bi-LSTM	Fake or Real News dataset	10342	94.26%	CIDII	51.49%
Bi-LSTM	CIDII	731	96.82%	CIDII	91.27%

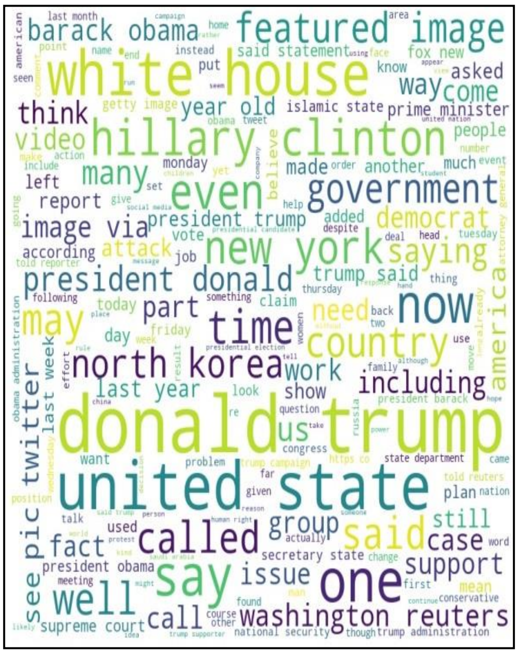


Fig. 27: The most frequent words in the Fake or Real News dataset

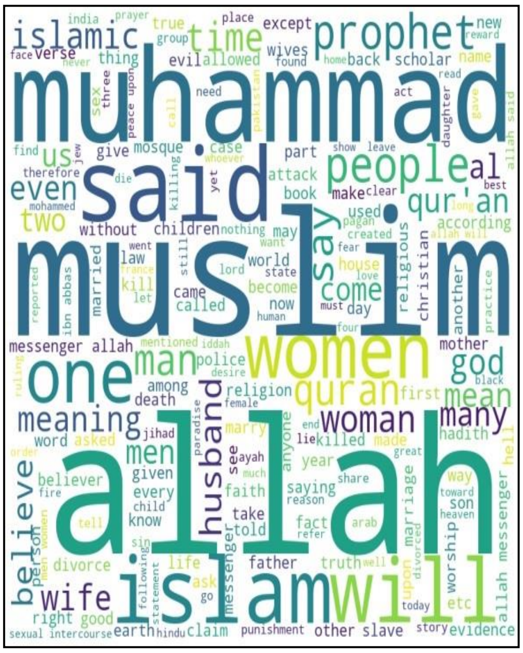


Fig. 28: The most frequent words in the CIDII dataset

Preprocessing may eliminate the notable features of disinformation through in-depth implementation. As we mentioned earlier in the research dataset analysis, most disinformation contains question marks, exclamation marks, or words written in capital letters; these attributes are not only in the CIDII dataset but in most standard datasets related to fake news [4], [39], [69]. These features, along with other salient features, are mainly used by the detection model to identify disinformation. According to the conducted experiments and as shown in Table 6, the selection of preprocessing steps that preserve the relevant features and eliminate noise contributed to an increase in the efficiency of the model and an increase in the AUC value of the results by 4.51 degrees over the performance of the same model but with strong preprocessing.

Table 7: Comparison of the model results if aggressive or non-aggressive preprocessing is used

Model	Preprocessing Type	Preprocessing Techniques	AUC Measure
Bi-LSTM	Aggressive Preprocessing	• Remove Links • Removing all special characters • Convert all letters to lowercase • Remove duplicate characters • Remove Accented Characters • Remove Whitespaces • Expand Contractions • Removing stopwords • Stemming	86.76
Bi-LSTM	Non-aggressive Preprocessing	• Remove Links • Remove Accented Characters • Remove duplicate characters • Expand Contractions • Removing special characters except (!,?) • Removing stopwords • Remove Whitespaces	91.27%

Moreover, the cased pre-trained GloVe model (300 D vectors, 2.03 GB) has contributed to enhancing the efficiency of the Bi-LSTM model by 1.36 degrees based on AUC measurement, compared to the Bi-LSTM model that uses the uncased pre-trained GloVe model (300 D vectors, 1.75 GB). This difference is due to the cased GloVe containing 840B tokens, including 2.2M vocabulary, which is rich in representations. It contains question marks, exclamation marks, and words in both uppercase and lowercase, while the uncased GloVe only includes lowercase words. Furthermore, it is larger than the uncased GloVe and contains 42B tokens, along with 1.9M words. In addition, the retrained GloVe model related to the specific domain, which was retrained on an Islamic Corpus, increased the efficiency of our proposed Bi-LSTM model and provided more accuracy than the rest of the models by 4.15 degrees according to the AUC measure as illustrated in Table 7. This result was due to the handling of OOV word problems, most of which are Islamic terms. Notably, the retrained GloVe for the Islamic domain was originally a cased pre-trained GloVe model (300 D vectors, 2.03 GB. The use of the cased GloVe model and then the use of the retrained GloVe model had a positive effect on the results and enhanced the performance of the different models used in these experiments, as shown in Table 7.

Table 8: Evaluation of Disinformation Baseline Models

Metric	**Pre-Trained GloVe Model Type**	**Model**			
		LSTM	**GRU**	**CNN**	**BI-LSTM (Our Proposed Model)**
AUC	300D Uncased GloVe Model	77.97%	86.45%	89.87%	89.91%
	300D Cased GloVe Model	80.12%	88.74%	90.65%	91.27%
	Retrained GloVe Model on Islamic Corpus	84.62%	90.81%	92.78%	95.42%
F1-Score	300D Uncased GloVe Model	83.56%	87.67	89.23%	91.73%
	300D Cased GloVe Model	85.14%	89.71%	91.93%	94.22%
	Retrained GloVe Model on Islamic Corpus	89.04%	91.41%	95.89%	97.31%

Figures 30 and 31 illustrate the training and validation AUC, and the training and validation loss of the proposed Bi-LSTM model.

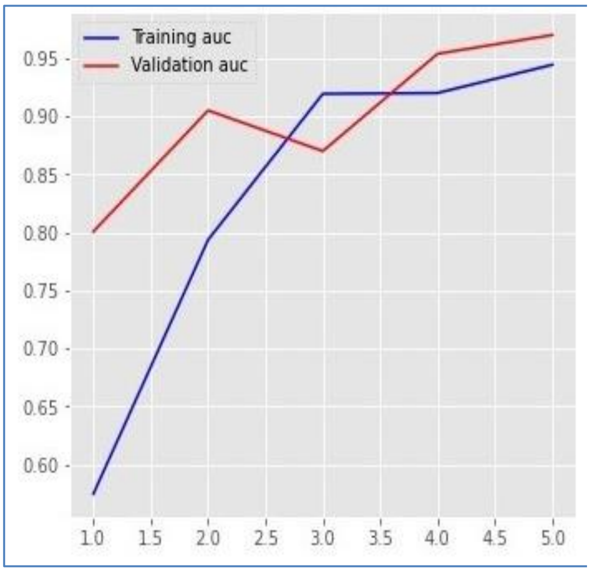


Fig. 29: The training and validation AUC of the proposed Bi-LSTM model

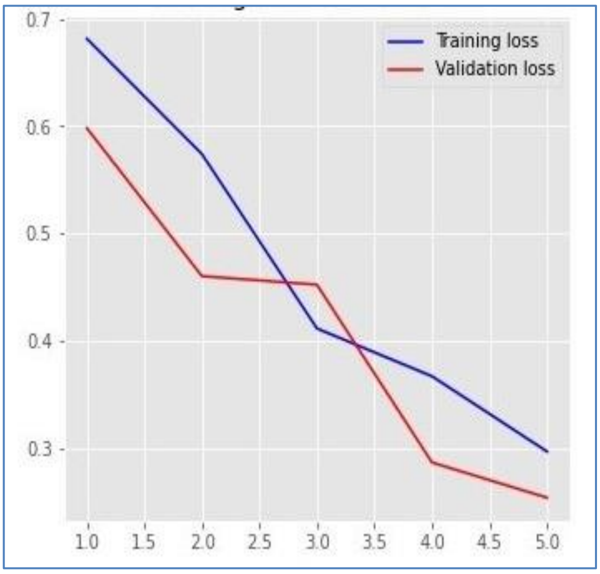
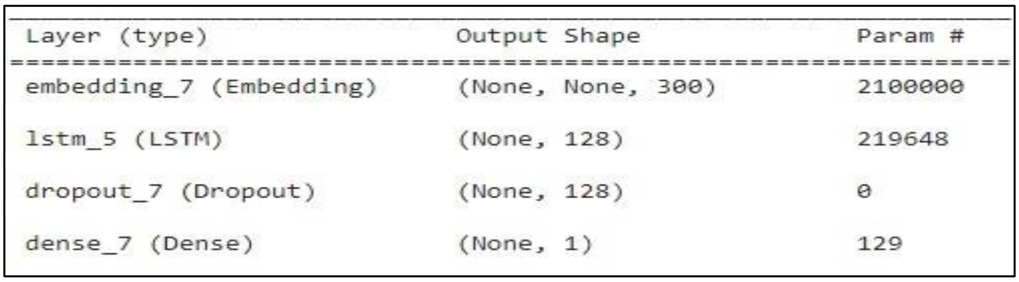


Fig. 30: The training and validation loss of the proposed Bi-LSTM model

To avoid under-fitting during the training of the used deep learning models due to the small size of the used dataset, we increased the complexity of the model as in the LSTM, GRU, and CNN, as illustrated in Figures 33, 34, and 35, or increased the number of epochs as in the Bi-LSTM model as shown in the previous Table 4.



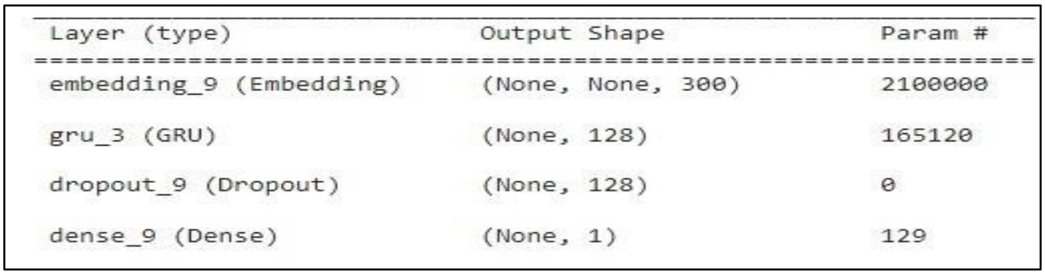
```
Layer (type)                 Output Shape              Param #
================================================================
embedding_7 (Embedding)      (None, None, 300)         2100000

lstm_5 (LSTM)                (None, 128)               219648

dropout_7 (Dropout)          (None, 128)               0

dense_7 (Dense)              (None, 1)                 129
```

Fig. 31: The configuration of the LSTM model



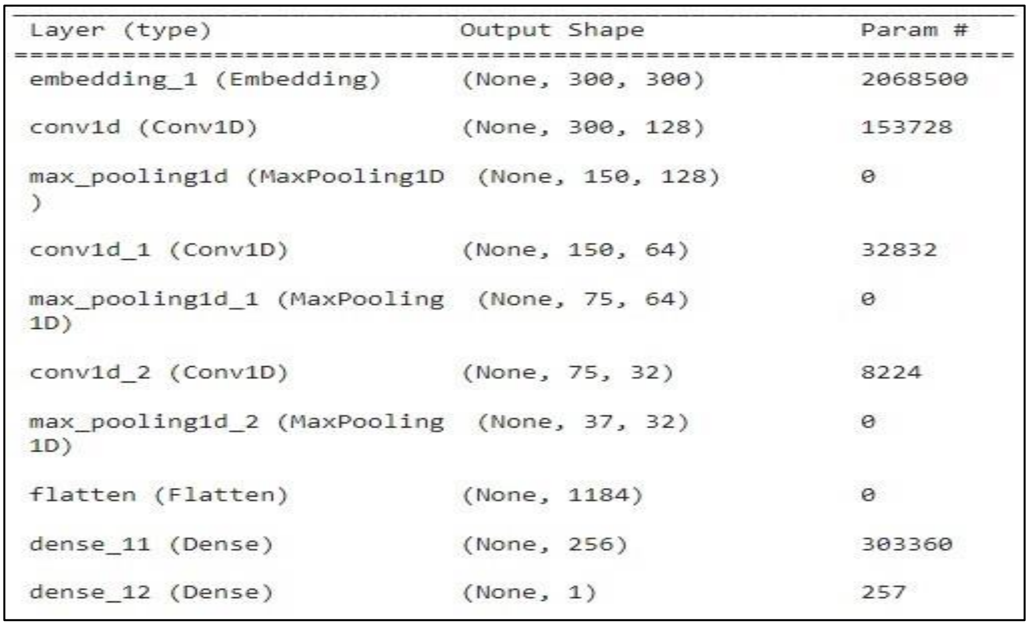
```
Layer (type)                 Output Shape              Param #
================================================================
embedding_9 (Embedding)      (None, None, 300)         2100000

gru_3 (GRU)                  (None, 128)               165120

dropout_9 (Dropout)          (None, 128)               0

dense_9 (Dense)              (None, 1)                 129
```

Fig. 32: The configuration of the GRU model



```
Layer (type)                 Output Shape              Param #
================================================================
embedding_1 (Embedding)      (None, 300, 300)          2068500

conv1d (Conv1D)              (None, 300, 128)          153728

max_pooling1d (MaxPooling1D   (None, 150, 128)         0
)

conv1d_1 (Conv1D)            (None, 150, 64)           32832

max_pooling1d_1 (MaxPooling   (None, 75, 64)           0
1D)

conv1d_2 (Conv1D)            (None, 75, 32)            8224

max_pooling1d_2 (MaxPooling   (None, 37, 32)           0
1D)

flatten (Flatten)            (None, 1184)              0

dense_11 (Dense)             (None, 256)               303360

dense_12 (Dense)             (None, 1)                 257
```

Fig. 33: The configuration of the CNN model

Based on the results presented in Table 7, we note that the most effective model that presented high detection accuracy according to the AUC metric is the Bi-LSTM model, with an accuracy of 95.42% of the AUC measure. This confirms the efficiency of the Bi-LSTM model in dealing with the text sequence; in addition, it is better than the unidirectional models.

6.0 CONCLUSION AND FUTURE WORK

The spread of disinformation against Islam on social media increases the spread of hatred against Muslims, reinforces the state of Islamophobia, and threatens societal peace. In this paper, we present a model for detecting disinformation about Islam based on the Bi-LSTM network. This model was trained based on an Islamic dataset (CIDII) collected by a team and labeled by a group of Muslim scholars. The accuracy of model detection was enhanced by retraining the GloVe word-embedding model on Islamic documents utilizing the Mittens method to address the problem of OOV words. This study showed that the model based on Bi-LSTM is better than the rest of the models in dealing with text sequences. Based on the results, the proposed model in this research and all the resources used can be utilized to combat the spread of anti-Islamic disinformation on social media. However, this research encountered difficulties in collecting data. For future works, we will address the limitation of the small size of the dataset using one of the data-augmentation techniques. In addition, we may explore new trends in the detection of disinformation using sentiment-based features or the public's stance if comments are available on such news in the dataset. Finally, instead of detecting disinformation, in the future, this research can be approached from another perspective and investigate “content debunking disinformation”.

ACKNOWLEDGEMENT

This research was funded by the Universiti Kebangsaan Malaysia under Geran Universiti Penyelidikan (GUP), Grant Code: GUP-2020-088.

REFERENCES

[1] Elhadad, M.K., 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)*. 2019. IEEE.

[2] Xu, K., et al., *Detecting fake news over online social media via domain reputations and content understanding.* 2019. **25**(1): p. 20-27.

[3] Ahmad, I.S., et al., *Beyond sentiment classification: A novel approach for utilizing social media data for business intelligence.* 2020. **11**(3).

[4] Eke, C.I., et al., *Sarcasm identification in textual data: systematic review, research challenges and open directions.* 2020. **53**(6): p. 4215-4258.

[5] Islam, M.R., et al., *Deep learning for misinformation detection on online social networks: a survey and new perspectives.* Soc Netw Anal Min, 2020. **10**(1): p. 82.

[6] Lin, L. and Z. Chen, *Social rumor detection based on multilayer transformer encoding blocks.* J Concurrency Computation: Practice Experience, 2021. **33**(6): p. e6083.

[7] Kapusta, J., et al., *Comparison of fake and real news based on morphological analysis.* 2020. **171**: p. 2285-2293.

[8] Aldayel, A. and W.J.P.o.t.A.o.H.-C.I. Magdy, *Your stance is exposed! analysing possible factors for stance detection on social media.* 2019. **3**(CSCW): p. 1-20.

[9] Kvetanová, Z., et al., *Debunking as a Method of Uncovering Disinformation and Fake News.* 2021.

[10] Vidgen, B., T.J.J.o.I.T. Yasseri, and Politics, *Detecting weak and strong Islamophobic hate speech on social media.* 2020. **17**(1): p. 66-78.

[11] Evolvi, G., *Hate in a tweet: Exploring internet-based islamophobic discourses.* J Religions, 2018. **9**(10): p. 307.

[12] Kaliyar, R.K., A. Goswami, and P.J.T.J.o.S. Narang, *DeepFakE: improving fake news detection using tensor decomposition-based deep neural network.* 2021. **77**(2): p. 1015-1037.

[13] Liu, Y. and Y.-F.B.J.A.T.o.I.S. Wu, *Fned: a deep network for fake news early detection on social media.* 2020. **38**(3): p. 1-33.

[14] Civila, S., L.M. Romero-Rodríguez, and A.J.P. Civila, *The Demonization of Islam through Social Media: A Case Study of# Stopislam in Instagram.* 2020. **8**(4): p. 52.

[15] D’Ulizia, A., et al., *Fake news detection: a survey of evaluation datasets.* 2021. **7**: p. e518.

[16] Janicka, M., M. Pszona, and A.J.C.y.S. Wawer, *Cross-domain failures of fake news detection.* 2019. **23**(3): p. 1089-1097.

[17] Shu, K., et al., *Mining disinformation and fake news: Concepts, methods, and recent advancements*, in *Disinformation, misinformation, and fake news in social media*. 2020, Springer. p. 1-19.

[18] Ahmad, I., et al., *Fake News Detection Using Machine Learning Ensemble Methods.* 2020. **2020**.

[19] Silva, A., et al. *Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data*. in *Proceedings of the AAAI conference on artificial intelligence*. 2021.

[20] Tashtoush, Y., et al., *A Deep Learning Framework for Detection of COVID-19 Fake News on Social Media Platforms.* 2022. **7**(5): p. 65.

[21] Gani, M.O., et al., *Bloom's Taxonomy-based exam question classification: The outcome of CNN and optimal pre-trained word embedding technique.* 2023: p. 1-22.

[22] Salur, M.U. and I.J.I.A. Aydin, *A novel hybrid deep learning model for sentiment classification.* 2020. **8**: p. 58080-58093.

[23] Li, J., S. Ni, and H.-Y.J.I.A. Kao, *Birds of a Feather Rumor Together? Exploring Homogeneity and Conversation Structure in Social Media for Rumor Detection.* 2020. **8**: p. 212865-212875.

[24] Al-Sarem, M., et al., *Deep learning-based rumor detection on microblogging platforms: a systematic review.* 2019. **7**: p. 152788-152812.

[25] Varshney, D. and D.K.J.J.o.A.I. Vishwakarma, *Hoax news-inspector: a real-time prediction of fake news using content resemblance over web search results for authenticating the credibility of news articles.* 2020: p. 1-14.

[26] Kumar, Y. and N.J.S.C.S. Goel, *AI-Based Learning Techniques for Sarcasm Detection of Social Media Tweets: State-of-the-Art Survey.* 2020. **1**(6): p. 1-14.

[27] Habib, A., et al., *False information detection in online content and its role in decision making: a systematic literature review.* 2019. **9**(1): p. 1-20.

[28] Farkas, J., et al., *Cloaked Facebook pages: Exploring fake Islamist propaganda in social media.* 2018. **20**(5): p. 1850-1867.

[29] Awan, I., *Islamophobia and Twitter: A typology of online hate against Muslims on social media.* 2014. **6**(2): p. 133-150.

[30] Balakrishnan, V., K.S. Ng, and H.A.J.T.i.S. Rahim, *To share or not to share–The underlying motives of sharing fake news amidst the COVID-19 pandemic in Malaysia.* 2021. **66**: p. 101676.

[31] Obadă, D.-R., D.-C.J.I.J.o.E.R. Dabija, and P. Health, *"In Flow"! Why Do Users Share Fake News about Environmentally Friendly Brands on Social Media?* 2022. **19**(8): p. 4861.

[32] Collins, B., et al., *Trends in combating fake news on social media–a survey.* 2021. **5**(2): p. 247-266.

[33] Mohamed, F.A., et al. *Identifying cues to deception in Islamic websites text-based content and design*. in *2018 International Conference on Information and Communication Technology for the Muslim World (ICT4M)*. 2018. IEEE.

[34] Al-Salemi, B., et al. *Feature selection based on supervised topic modeling for boosting-based multi-label text categorization*. in *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI)*. 2017. IEEE.

[35] Hajek, P., et al., *Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining.* 2020. **32**(23): p. 17259-17274.

[36] Shu, K., et al., *Fake news detection on social media: A data mining perspective.* 2017. **19**(1): p. 22-36.

[37] Wang, L., et al., *Understanding archetypes of fake news via fine-grained classification.* 2019. **9**(1): p. 1-17.

[38] Ghosh, S., C.J.P.o.t.A.f.I.S. Shah, and Technology, *Towards automatic fake news classification.* 2018. **55**(1): p. 805-807.

[39] Zhou, X., et al., *Fake news early detection: A theory-driven model.* 2020. **1**(2): p. 1-25.

[40] Aldwairi, M. and A.J.P.C.S. Alwahedi, *Detecting fake news in social media networks.* 2018. **141**: p. 215-222.

[41] Kumar, S., et al., *Fake news detection using deep learning models: A novel approach.* 2020. **31**(2): p. e3767.

[42] de Oliveira, N.R., D.S. Medeiros, and D.M.J.I.S.P.L. Mattos, *A sensitive stylistic approach to identify fake news on social networking.* 2020. **27**: p. 1250-1254.

[43] Kaur, S., P. Kumar, and P.J.S.C. Kumaraguru, *Automating fake news detection system using multi-level voting model.* 2020. **24**(12): p. 9049-9069.

[44] Vicari, M., M.J.A. Gaspari, and Society, *Analysis of news sentiments using natural language processing and deep learning.* 2020: p. 1-7.

[45] Abdelminaam, D.S., et al., *CoAID-DEEP: An Optimized Intelligent Framework for Automated Detecting COVID-19 Misleading Information on Twitter.* IEEE Access, 2021. **9**: p. 27840-27867.

[46] Liang, H., et al., *Text feature extraction based on deep learning: a review.* EURASIP J Wirel Commun Netw, 2017. **2017**(1): p. 211.

[47] Kumar, A., et al., *Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network.* 2019. **7**: p. 23319-23328.

[48] Deepak, S. and B.J.P.C.S. Chitturi, *Deep neural approach to Fake-News identification.* 2020. **167**: p. 2236-2243.

[49] Subramani, S., et al., *Domestic violence crisis identification from facebook posts based on deep learning.* 2018. **6**: p. 54075-54085.

[50] Bahad, P., P. Saxena, and R.J.P.C.S. Kamal, *Fake news detection using bi-directional LSTM-recurrent neural network.* 2019. **165**: p. 74-82.

[51] Asghar, M.Z., et al., *Exploring deep neural networks for rumor detection.* 2021. **12**(4): p. 4315-4333.

[52] Barbado, R., et al., *A framework for fake review detection in online consumer electronics retailers.* 2019. **56**(4): p. 1234-1244.

[53] Elhadad, M.K., K.F. Li, and F. Gebali, *Detecting Misleading Information on COVID-19.* IEEE Access, 2020. **8**: p. 165201-165215.

[54] Mridha, M.F., et al., *A Comprehensive Review on Fake News Detection with Deep Learning.* 2021.

[55] Balakrishnan, V., H.L. Zing, and E.J.M.J.o.C.S. Laporte, *COVID-19 INFODEMIC–UNDERSTANDING CONTENT FEATURES IN DETECTING FAKE NEWS USING A MACHINE LEARNING APPROACH.* 2023. **36**(1): p. 1-13.

[56] Hamed, S.K. and M.J.J.J.C.S. Ab Aziz, *A Question Answering System on Holy Quran Translation Based on Question Expansion Technique and Neural Network Classification.* 2016. **12**(3): p. 169-177.

[57] Thalib, P., *Distinction of characteristics sharia and fiqh on islamic law.* Yuridika, 2018. **33**(3): p. 439-452.

[58] Ahmad, S.R., A.A. Bakar, and M.R.J.I.d.a. Yaakub, *A review of feature selection techniques in sentiment analysis.* 2019. **23**(1): p. 159-189.

[59] Saif, A., et al., *Semantic concept model using Wikipedia semantic features.* 2018. **44**(4): p. 526-551.

[60] Pennington, J., R. Socher, and C.D. Manning. *Glove: Global vectors for word representation*. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

[61] Batbaatar, E., M. Li, and K.H.J.I.A. Ryu, *Semantic-emotion neural network for emotion recognition from text.* 2019. **7**: p. 111866-111878.

[62] Kaliyar, R.K., A. Goswami, and P. Narang, *FakeBERT: Fake news detection in social media with a BERT-based deep learning approach.* Multimed Tools Appl, 2021. **80**(8): p. 11765-11788.

[63] Ilie, V.-I., et al., *Context-Aware Misinformation Detection: A Benchmark of Deep Learning Architectures Using Word Embeddings.* 2021. **9**: p. 162122-162146.

[64] Kilgarriff, A., et al., *The Sketch Engine: ten years on.* 2014. **1**(1): p. 7-36.

[65] Dingwall, N. and C.J.a.p.a. Potts, *Mittens: an extension of glove for learning domain-specialized representations.* 2018.

[66] Wang, J.-H., T.-W. Liu, and X.J.A.S. Luo, *Combining Post Sentiments and User Participation for Extracting Public Stances from Twitter.* 2020. **10**(22): p. 8035.

[67] de Souza, J.V., et al., *A systematic mapping on automatic classification of fake news in social media.* 2020. **10**(1): p. 1-21.

[68] Subramani, S., et al., *Deep learning for multi-class identification from domestic violence online posts.* 2019. **7**: p. 46210-46224.

[69] Faustini, P.H.A. and T.F.J.E.S.w.A. Covões, *Fake news detection in multiple platforms and languages.* 2020. **158**: p. 113503.

APPENDIX I

ISLAMIC DOMAIN EXPERTS

No.	Name	Degree	University	Specialty	Position	Languages	Email
1	Laith Salman Dawood	Ph.D.	Baghdad	Islamic Creed and Thought	Imam and preacher	Arabic, English	dr.laithsalamn@gmail.com
2	Diar Mahmood Saeed	Ph.D.	USIM	Quran Interpretation	Lecturer at Islamic University of Minnesota	Arabic, English	diarmirany@gmail.com
3	Zayd Thabit Abdul Rahman	Ph.D.	USIM	Philosophy of Quran and Sunnah Studies	Imam and preacher	Arabic	dr.zaid1@yahoo.com

APPENDIX II

DATASET COLLECTORS TEAM

No	Name	Degree	University	Specialty	Languages	Email
1	Suhaib Kh. Hamed	Ph.D. candidate	UKM	Artificial Intelligence	Arabic, English	p105401@siswa.ukm.edu.my
2	Othman Arif Hanshal	Ph.D.	Altinbas	Artificial Intelligence	Arabic, English	p90703@siswa.ukm.edu.my
3	Saad Mahmoud Ahmed	Ph.D. candidate	UKM	Information System	Arabic, English	othman.hanshal@ogr.altinbas.edu.tr