Bulletin of Electrical Engineering and Informatics
Vol. 13, No. 6, December 2024, pp. 4260~4272
ISSN: 2302-9285, DOI: 10.11591/eei.v13i6.8016

Machine learning-based detection of fake news in Afan Oromo language

Ayodeji Olalekan Salau[1,3,7], Kedir Lemma Arega[2,8], Ting Tin Tin[3], Andrew Quansah[4], Kwame Sefa-Boateng[5], Ismatul Jannat Chowdhury[4], Sepiribo Lucky Braide[6]
[1]Department of Electrical/Electronics and Computer Engineering, Afe Babalola University, Ado-Ekiti, Nigeria
[2]Department of Information Technology, School of Technology and Informatics, Ambo University, Oromia, Ethiopia
[3]Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia
[4]Department of Electrical and Computer Engineering, University of North Carolina, Charlotte, USA
[5]College of Computing and Informatics, University of North Carolina at Charlotte, North Carolina, USA
[6]Department of Electrical and Electronics Engineering, Rivers State University, Port Harcourt, Nigeria
[7]Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India
[8]Department of Software Engineering, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia

Article Info

Article history:

Received Dec 11, 2023
Revised May 4, 2024
Accepted May 17, 2024

Keywords:

Afan Oromo
Classification
Detection
Ethnic conflict
Fake news
Machine learning

ABSTRACT

This paper presents a machine learning-based (ML) approach for identifying fake news on web-based social media networks. Data was acquired from Facebook to develop the model which was used to identify Afan Oromo's false news. The system architecture uses algorithms, such as support vector machines (SVM), k-nearest neighbor (KNN), and convolutional neural networks (CNNs) to detect and classify fake news. Existing models have limitations in understanding reported news accuracy compared with verified news. This study successfully resolved the challenges in the detection of social media fake news detection for the Afan Oromo language with the use of ML models and natural language processing (NLP) techniques. The results show that the SVM approach achieved a precision, recall, and F1-score, of 0.92, 0.92, and 0.90.

Corresponding Author:

Ayodeji Olalekan Salau
Department of Electrical/Electronics and Computer Engineering, Afe Babalola University
Ado-Ekiti, Nigeria
Email: ayodejisalau98@gmail.com

1. INTRODUCTION

Online social networks are a popular means of communication. Numerous persons engage in interpersonal interactions and disseminate information daily through visual content as well as post updates on their present status on social media platforms. Many individuals possess a fondness for employing social networking platforms, such as Facebook because of the rapidity with which information can be shared and friendly user interface [1]. With the development of communication technologies and social media, fake news has rapidly expanded. Fake news identification is a recent research area that has gained a lot of interest [2]. Fake news is a major issue in today's social media and political landscape. The detection of fake news necessitates extensive study, although there are several challenges [3]. Misinformation is a serious problem involving various individuals that disseminate content that might be truthful or damaging. The spread of false news is a deliberate act, and the use of online social media networks by the users presents unique challenges. Traditional news sources' identification algorithms are inadequate, but people seek news from social networks due to their ease of access, minimal effort, and rapid dissemination [4], [5]. Social media is a popular medium for individuals to share their thoughts, views, and news. Afan Oromo, Africa's largest language, is widely spoken in Ethiopia and neighboring nations. It provides fiction, literature, and journalism, as well as the ability for spectators and witnesses to speak about occurrences. This media has a tremendous influence on people's views and beliefs. Because of the differences in language, it is difficult to discern between fake and actual news [6]. Fake news or deceptive news, has existed prior to the invention of computers and the internet. It might include spreading rumors about public or private events, casually discussing people's lives, or harming competing companies' reputations.

People nowadays expect rapid, low-cost information broadcasts which has made internet content to gain significant traction. Social media has outperformed traditional news transmission systems in terms of decision-making and opinion formation [7]. Aside from traditional news sources, online social networks serve as the key platform for disseminating news and user perspectives. The explanation for this shift in user behavior on social media platforms is that information distribution on social media is typically more consistent and economically advantageous [8].

In recent years, the rise of deceptive content, such as the spread of fake news and misleading information, has created an increasing risk to those who use social networking sites like Facebook, Twitter, and Google Plus. Scholarly research has indicated that an incredible amount of postings, exceeding one trillion per second, are continuously created on the Internet, mostly via social media platforms such as Facebook. As a result, the usual process of checking factual correctness of the information disseminated on such platforms faces significant hurdles [9]. Initiatives to combat fake news have emerged as its spread has reached critical levels [10]. This study investigates the detection of fake news in Afan Oromo language which has rich morphology, focusing on the use of deep learning algorithms. Despite widespread use in Ethiopia, earlier research on fake news identification on social media platforms such as Facebook is insufficient. The work analyzes features of fake news and presents classification datasets, with the goal of improving fake news detection and classification in Afan Oromo language. Ethiopian social media sites have profited from the use of machine learning (ML) to detect fake news, but they may also propagate false information [11]. In 2018, the government's grip over social media diminished, resulting in more freedom of speech. However, the COVID-19 epidemic has heightened the dangers of hate speech and fake news, complicated response efforts and making it difficult for medical professionals to respond correctly [12]. To mitigate the negative impacts, governments, the technology industry, and researchers are suggesting novel strategies for combating false news transmission [13], [14]. In Ethiopia and other African countries, Facebook uses an independent verification expert service, pursuant to the "Hate speech and disinformation prevention and suppression proclamation No. 1185/2020" [15]. However, this assertion is problematic in terms of equality and freedom of expression [11]. Ethiopian laws aimed at preventing hate speech and disinformation dissemination have been challenged by fake news creators in recent times [16]. Despite the addition of four new working languages, Ethiopian spoken languages remain among the world's most threatened "low resources" due to a scarcity of resources for natural language processing (NLP). Research has been undertaken in Ethiopia and other foreign languages to detect and combat fake news in internet interactions. The research's goal is to create a false news detection system for Afan Oromo, assess its performance, and report on the findings. Based on performance and accuracy tests, the best detection model for Afan Oromo false news is recommended.

2. AFAN OROMO LANGUAGE

2.1. Overview of the Afan Oromo language

According to Debela [17], the Oromo people and other nearby ethnic groups in the Horn of Africa speak Afan Oromo, an Afro-Asiatic language that is a member of the Cushitic branch. There were 33.8% Afan Oromo speakers, followed by 29.3% Amharic language speakers. In Ethiopia, Afan Oromo is commonly spoken and written. As for its writing system, Qubee (a Latin-based alphabet) has been used as Afan Oromo's official script since 1991 [17].

2.2. Afan Oromo Qubee and writing system

Qubee Afan Oromo is the name of the Afan Oromo language's alphabet, which uses capital and lowercase letters similar to the English alphabet [18]. The vowels and consonants of the Afan Oromo language are identical to those in English. Vowels in the Afan Oromo language are represented by the five fundamental letters "a", "e", "i", "o", and "u." Additionally, it has the standard set of five short and five long vowels found in Eastern Cushitic by duplicating the five vowel letters: "aa", "ee", "ii", "oo", and "uu" [19]. Consonants in the Afan Oromo language, on the other hand, are not significantly different from English letters in written texts, except for a few special combinations known as "Qubee Dachaa" such as "sh" and "ch" (which sound similar to English), and "dh" (which sounds similar to an English "d") produced with the tongue slightly curled back and the air drawn in so that a glottal stop is heard before the next vowel begins.

Another combination is "ph", which is produced when "ny", when smacked with the outside of the mouth, and closely resembles the English sound of "gn". These few unique letter combinations are frequently used to create words. For instance, the letters “dh is used in the word "butter," "ch" is used in the word "important", "sh" is used in the word "girl", "ph" is used in the word "egg" and "ny" is used in the word "food". The word "Qubee" is made up of 36 letters in the Afan Oromo language (26 consonants and 10 vowels). Similar to how it is done in English, white space is utilized to separate words in phrases. The various punctuation marks used in Afan Oromo follow the same pattern as those used in English and other languages that employ the Latin writing system. For instance, the full stop (.) in a statement, the question mark (?) in an interrogative sentence, and the exclamation mark (!) in a command or exclamatory sentence all serve to denote the end of a sentence [17]. In general, every letter in the English alphabet has a counterpart in Afan Oromo; except for the manner in, which it is written. Using Afan Oromo dialects in the Afan Oromo language, vowels can exist in the start, medial, and final positions of a word. Everywhere that a short vowel can appear, a long vowel is heard as a single unit.

2.3. Afan Oromo punctuation mark

Punctuation is used in language structure to clarify meaning and facilitate reading. Different punctuation marks used in Afan Oromo's writings are found to follow the same punctuation pattern as those used in English and other languages that use the Latin writing system [19]. The following are some of the most often used punctuation marks in Afan Oromo, which uses the same text structure as the English language: i) tuqa full stop (.) is used at the end of a sentence and in abbreviations; ii) mallattoo gaffii question mark (?) is used in inquiring or at the end of a straight question’; iii) rajeffannoo exclamation mark (!) is used at the end of knowledge and exclamatory sentences; iv) qooddu comma (,) is used to isolate listing in a sentence or to separate the elements in a sequence; and v) tuq-lamee colon (:) is the purpose of the colon is to separate and present lists, sections, and quotations, along with several conventional uses.

2.4. Word structure

The fundamental building block of a language is called a "jecha" in Afan Oromo. Additionally, it refers to a group of sounds that can stand on their own and convey meaning [20]. Claims that Afan Oromo words can range from a small number of monosyllabic words to polysyllabic words with up to seven syllables. The language has an easy-to-understand writing system, which implies that everything is written as it is read and vice versa. White space characters are also used to distinguish Afan Oromo words from one another. Therefore, the process of taking an input text and adding valid word boundaries, known as word segmentation (tokenization), is carried out by utilizing the white space characters for information retrieval purposes [20]. The smallest linguistic unit is the word. Words can be separated from one another using a variety of techniques. However, the majority of languages, including English, utilize the space symbol to indicate the end of a word. The symbols "/" and "." are used to break up some long words when they are written (abbreviated), thus they shouldn't serve as a word boundary. In Afan Oromo, word boundaries are indicated using the standard parenthesis, brackets, quotes, and other punctuation [21].

2.5. Sentence construction

English and Afan Oromo have different sentence structures and syntax. The subject-object-verb (SOV) structure is used in the Afan Oromo language because its word order is flexible or varied. A Sentence building known as subject-verb-object (SVO) places the subject first the verb second and the object third. For instance, "Mohammad barista dha" or "barsisaa dha Mohammad" are preferred in the Afan Oromo language. A subject is "Mohammad", an object is "Barsiisa", and a verb is "dha". It has an SOV structure as a result. "Mohammad is a teacher", which is how the sentence is translated in English, has a set word order and an SVO structure. Additionally, the way that adjectives are formed in English and Afan Oromo differs. While English adjectives often precede the noun, Afan Oromo adjectives agree with their head noun, which is not the case in English. In Afan Oromo, adjectives follow a noun or pronoun; their customary position is near the word they modify. As an illustration, miicaayyoo bareeduu (stunning girl) comes after miicaayyoo (noun) [20]. Similar to English and other languages that use the Latin writing system, Afan Oromo sentences are finished. That is to say the full stop (.) that means the comma (,), which separates lists in a sentence, and the semicolon (;) are to mark a break that is stronger than a comma but not as final as a full stop balance. The full stop (.) in a statement, the question mark (?) in interrogative, the exclamation mark (!) in command and exclamatory sentences, and the exclamation mark (!) in command and exclamatory sentences mark the end of a sentence [21].

3. RELATED WORK

Various scholars have utilized techniques that analyze the occurrence, utilization, and patterns in written content to detect inaccurate and deceitful materials. By doing so, they can pinpoint similarities that conform to established textual conventions, such as misleading news, which employs language resembling satire, exhibits a higher degree of emotional language, and is less complex than articles about the same topic [22]. Artificial intelligence (AI) and ML techniques have been incorporated to improve the effectiveness of linguistic systems. The popularity of social media online news is growing as a result of the internet's quick development. Fake news is also frequently produced and is spreading to deceive readers. Therefore, study into fake news has been intense in recent years to combat the issue of fake news on online social media. The researchers in [6], [21], [23] suggest analyzing news textual data using text mining and ML approaches to forecast the news's reliability.

To comprehend the idea behind this study, a thorough examination of similar ideas and earlier research in this field was conducted [2]. Provide a strategy for identifying fake news that makes use of ML methods. As a feature extraction strategy, they employed the term frequency-inverse document frequency (TF-IDF) of a sample of words and n-grams, and support vector machine (SVM) as a classifier. The authors also suggest a dataset of real and fake news for the suggested system's training. Results obtained demonstrate the system's effectiveness. Through the use of various categorization algorithms by Shaikh and Patil [24] work enables us to identify fake news that is accurate. Research by Shaikh and Patil [24], social lives are being negatively impacted by fake news, which also has a major impact on politics and education. The authors used classification methods like SVM, naive Bayes (NB), and passive aggressive (PA) classifier in this model. The accuracy of the results produced by their model, which uses feature extraction methods like TF-IDF and SVM as classifier, is 95.05%. By suggesting a system that can accurately categorize fake news, this research aims to hasten the process of fake news identification [25]. On eight different datasets obtained from various sources, ML algorithms such as NB, PA classifier, and deep neural networks (DNN) have been applied. The document also features each model's analysis and outcomes [26]. In this paper the research used ten different ML and deep learning classifiers were used to categorize the fake news dataset in this classification. Four traditional methods (TF-IDF, count vector, character level vector, and n-gram level vector) were used to extract features from texts.

The outcomes demonstrated that fake news with textual content may be identified, particularly when convolutional neural networks (CNNs) are used. Using various classifiers, this study obtained an accuracy range of 81% to 100%. The author suggested using the three well-known ML methods NB, neural network, and SVM in this proposed study [27]. The accuracy of the NB result was 96.08%, while the accuracy of the two other, more advanced approaches, the support vector and neural network, was 99.90%.

In this study, the researchers present an automatic fake news detection system that supports or refutes questionable assertions while returning a set of documents from trusted sources. The system is made up of several modules and employs techniques from ML, deep learning, and NLP. Such strategies are used to select relevant papers and identify those that are comparable to the tested claim and their viewpoints. The suggested approach will be used to validate medical news, particularly those on the COVID-19 epidemic, vaccine, and cure [28].

Ethiopian regional language fake news detection research is currently in its early stages [29]. Describes the use of multinomial NB with NLP algorithms for the identification of bogus news in 752 Afan Oromo news texts [23]. They applied the TF-IDF of unigrams and bi-grams and used Facebook as the news article's source. Based on their data size, they were able to obtain good outcomes with the best F1 score. Although the accuracy acquired is good, the information from the data set site's source is not entirely reliable, and the highest accuracy might have been attained by considering unreliable news. Spam emails and fake news pieces undoubtedly share several crucial characteristics. Therefore, using similar techniques for spam filtering and fake news detection is crucial [30].

Offer a model for detecting bogus news that makes use of n-gram analysis and ML. Other studies, such as that by Bajaj [31], focused on the issue of fake news. To develop a classifier that can determine if a piece of news is real or fake based on the news's content and compare the outcomes of various models, the authors used pre-trained 300-dimensional glove embeddings. Further along the line, neural network design is shown to forecast the stance between a given pair of headlines and the article body. This architecture outperforms existing model architectures and achieves a higher F1 score, filling the gap left by binary classification [32].

As a result, deep learning models perform better than ML methods [29]. Ganfure [33] utilized feature-based techniques to rate the trustworthiness of tweets on Twitter and succeeded to some extent, but the study mainly relied on feature engineering, which is costly and time-consuming. As a result, more recent deep learning projects were carried out to eliminate the requirement for feature engineering.

Twitter streams were modeled as sequential data and then a recurrent neural network (RNN) was utilized to determine whether or not the streams were phony [34]. These research studies were carried out on one of the most prominent datasets, the Koirala dataset. Based on the results of the prediction, the suggested model demonstrated an optimistic and superior predictability performance with high accuracy (75.4%) and reduced the number of features to 303. Furthermore, when compared to other state-of-the-art classifiers, their findings revealed that the suggested detection approach with the genetic algorithm model beat other classifiers in terms of accuracy [33]–[37].

This study examines the phenomenon of fabricated news in Ethiopia, with a particular focus on the significance of social media literacy. Approaches to categorizing misinformation encompass linguistic, visual, social context, and interaction evidence. ML, bag-of-words, and predictive models are employed to extract features from textual data [38]. Various authors have devised frameworks for various languages to capture explicit or implicit attributes, thus enabling models to draw generalizations. The investigation also delves into the identification of false information in other languages, such as Afan Oromo. Nonetheless, further research is required to formulate effective methodologies for detecting false information in these languages. Different approaches for fake news detection are depicted in Figure 1 and the notations used in this paper are presented in Table 1.

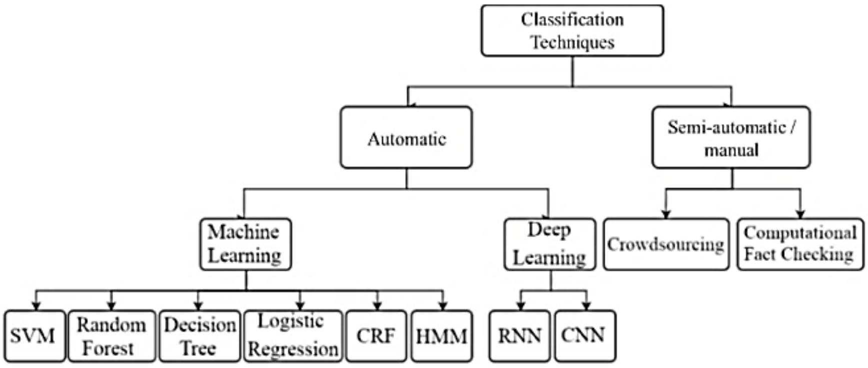

Figure 1. Approaches of fake news detection

Table 1. Summarize the notations used in the paper

Symbol	Descriptions
1185/2020	In the Ethiopian case, 'hate speech and disinformation prevention and suppression proclamation, No. 1185/2020' has been enacted. The proclamation indicates prohibited acts of hate speech and its exceptions
No.	Number
N	Refers to the number of samples in the training dataset
SOV	Subject-object-verb
RNN	Recurrent neural network
TF-IDF	Term frequency-inverse document frequency
SVM	Support vector machine
KNN	K-nearest neighbor
NN	Neural network
TP	True positives
FP	False positives
FN	False negatives
TN	True negatives
LSTM	Long short-term memory
BiLSTM	Bidirectional long short-term memory
Ch, Dh, Ny, Ph, and Sh	Special double letters include Ch, Dh, Ny, Ph, and Sh. These double letters are known as "Qubee dacha" in Afaan Oromo

4. METHOD
4.1. Proposed system architecture

The notion of fake news identification is a developing research area in the field of computer science, and there is a substantial body of literature on the subject for texts and datasets that are based in English. Fake news identification in Afan Oromo is still in its infancy, despite the current surge in interest in the subject. To train the classifiers, however, there were no datasets for Afan Oromo text until recently. The design, development, and classification of Afan Oromo false information distributed via social media are covered in this phase. Social media is a platform where anybody can share their positive thoughts and form opinions. On the other hand, fake news, phony reviews, and bad comments have grown to be risky for online users. The main difficulty in developing algorithms for Afan Oromo text false news identification, in particular, is finding a large, rich, and consistently labeled dataset. Additionally, the definition of "fake news" is fairly ambiguous and complicated [10].

The primary source of data for this study was Facebook. The system architecture is an abstract representation of a system's structure, behavior, and other aspects. Since assigning tags or categories to texts following their content is a sort of text classification activity, false news detection falls under this category. The datasets collected, cleaned, and partitioned; feature engineering; selecting the appropriate algorithm; mathematical models and methodologies; and results make up the suggested system architecture for detection and classification of Afan Oromo fake news that is shared on social media shown in Figure 2.

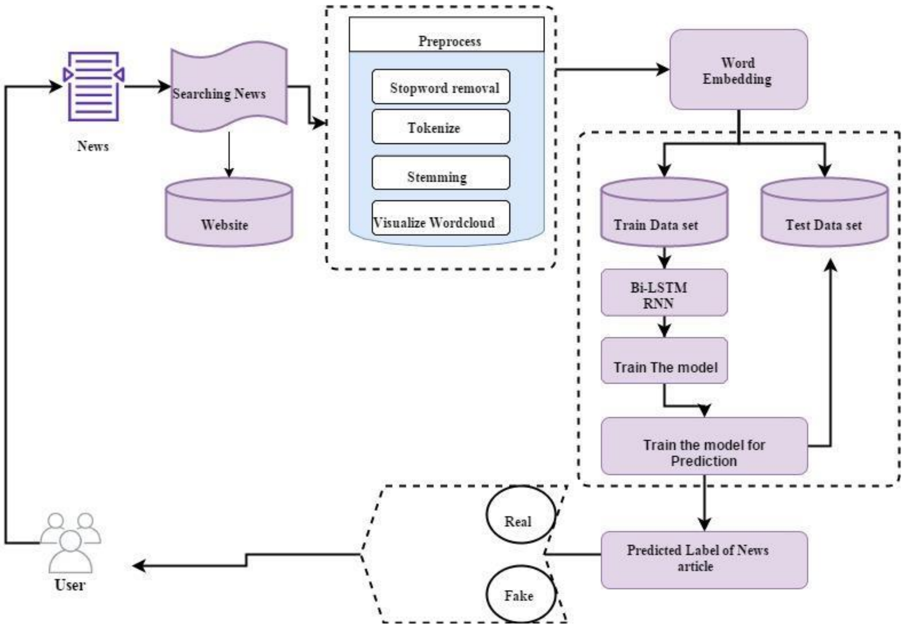

Figure 2. Proposed system architecture for identifying and categorizing Afan Oromo fake news disseminated via social media

4.2. Dataset gathering and corpus construction

The initial step in developing our models was to obtain the data that we would be working with. Social media has evolved into a platform for people to express their interests, opinions, and displeasure, among other things. As a result, examining this data was a natural fit for our study. For this study to be completed quickly, knowing the total number of Afan Oromo posts made on Facebook from day to day is ideal, but it's challenging to gather these numbers. Because of this, the study gathered datasets of fake news articles from Facebook with the help of subject-matter experts for experimentation. Up till now, Afan Oromo has not had publicly accessible assembled datasets that were used for tasks like fake news identification and categorization on social media. As a result, the news dataset used in this study's experiment was manually compiled by a subject matter expert from several Facebook pages. Since there has been insufficient prior research or corpus used to evaluate Afan Oromo's fake news detection and classification, the corpus used for this evaluation (Afan Oromo news posts) was manually compiled. The collected and ready corpus was then preprocessed through a variety of procedures, including tokenization and normalization.

4.3. Data preprocessing

The study focuses on the detection of fake news in Afan Oromo using two datasets: fake and real news. The data is prepared through a process that includes gathering, cleaning, filtering, and consolidating it into a single file or table. The data is then used for annotation of user activities and timelines, followed by the creation of a training model for fake news detection. Two alternative approaches are used, both before and after preprocessing, to extract feature engineering such as capital letters, punctuation marks, and news article/document durations. The data is cleaned for further analysis using dataset preparation. Text categorization is a complex process, requiring multiple phases and back-end effort. The algorithms compare news based on factors such as true and false examples, accessibility of digital formats, verifiability of ground truth, homogeneity of lengths, writing subjects, predetermined timeframes, and delivery method. The study used Facebook datasets, which were prepared in the format presented in Table 2. Although the statements in the dataset are news articles that are thought to be on various community topics, social, economic, technological, and political issues, the researcher has to deal with a two-label setting (i.e., true, fake), so they are a potential source for gathering a balanced corpus for the task of detecting fake news for the Afan Oromo Language. To represent texts for the most basic aspect of deep learning, we sum up the main strategies in Figure 3.

Table 2. Dataset preparation format

Feature	Category
Heading	Text
Label	Number
Content	Text

	MataDuree	Qaama	Label
0	Barak Husen Obaman kitabaa barreessan\n\nBaara...	Pireeziantiin Ameerikaa duranii Baaraak Obaama...	1
1	Hidhamtoota siyaasa itiyoophiyaa	Obbo Iskindir Naggaa fi Sintaayyoo Chakool man...	1
2	Ergaan ykn barreeffamni isin facebook irratti ...	Dhaabbanni Facebook wantoota barreessitan isin...	1
3	Qonnaan bultoonni sababa misoomaatiin lafa isa...	Dhimmaa Qonnaan bultoota naannoo Finfinnee	1
4	Jijjirama siyaasa itiyoophiyaa\n\nSiyaasni Oro...	A.L.I barri 2012 xumurame taateewwan siyaasa s...	1
...	...	...	...
7186	Bareessaan olaanaan kun wayita wal ga'iin dhaa...	NaN	0
7187	Yeroo dhihoo asitti kutaa Sahaalitti walitti b...	NaN	0
7188	Humni farra shororkeessummaa naannichaa J-5 je...	NaN	0
7189	Biyyoonni kutaa Sahaal shaman shororkeessummaa...	NaN	0
7190	Torbee kana gamtaan Awurooppaa kutaa Sahaalitt...	NaN	0

7191 rows × 3 columns

Figure 3. Claims and ancillary data in the dataset

4.4. Feature extraction

The creation or extraction of features from the dataset is known as feature extraction [35]. The process of converting textual data into real-valued vectors is known as feature extraction. For many jobs in NLP, documents are frequently represented by linguistic attributes. The linguistics process refers to characteristics that show how text functions, like the typical number of words per phrase and the frequency of misspellings. Lexical features, such as character-level and word-level features like total words, characters per word, frequency of large words, and unique words; syntactic features, such as sentence-level features like frequency of function words and phrases like n-grams and bag-of-words approaches; and morphological features are typical examples of common linguistic features [36]. The process of feature extraction involves turning random input, such as text or photos, into numerical features that can be used in ML. In the latter, these traits are subjected to a machine-learning technique (https://scikit-learn.org, 2007-2019) [37]. In this study, the researcher proposes a system that identifies and detects Afan Oromo fake news broadcasts on social media using phrase frequency and inverse document frequency, together with an n-gram model. News content features define the Metadata associated with a piece of news. The following are some examples of representative news content attributes, the author or publisher of the news piece is the source: i) headline is brief title text that seeks to capture readers' attention and describes the primary topic of the content; ii) body text is the main text elaborates on the details of the news article; normally, there is a key allegation that is especially highlighted and shapes the publisher's stance; and iii) image or video is section of a news article's body material that provides visual clues to frame the story. Several types of feature representations can be created based on these raw content qualities to extract discriminative aspects of fake news.

4.5. Data visualization

To find the links between mountains of unstructured data and turn the unseen data into a visible graph, which aids the reader in rapidly identifying significant points, data visualization was utilized in the false news detection process. By adjusting the size of each word proportionate to its frequency and arranging the words into a cluster or cloud of words, the researcher can view how frequently words appear in a given text. Word clouds for the fake and real texts are shown in Figure 4 [7], while Figure 5 presents the term cloud for the script that is real. Words with multiple ascenders and descenders in their letters may be given greater attention, and long words are prioritized over short ones.

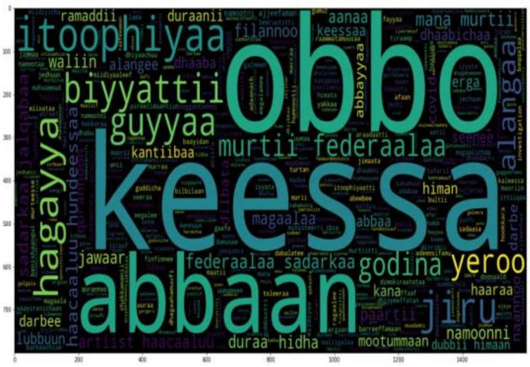

Figure 4. Word cloud for authentic text

Figure 5. Term cloud for the script that is real

5. RESULTS

Various evaluation measures will be utilized to assess the effectiveness of algorithms for the false news identification problem. The majority of current methods view the fake news problem as a classification issue that seeks to determine if a news story is false or not [36]. The Afan Oromo false news dataset was used in this study to test the performance of the ML models and the confusion matrix was used to represent the classification's four possible outcomes. In this study instance: i) true positives (TP) is the news item has been deemed false and is fake; ii) false positives (FP) is the news item has been mistakenly identified as true despite being false; and iii) false negatives (FN) is the news item is accurate but has been labeled as phony. The news item has been evaluated as true and is true, according to the true negative (TN). The authors extract the following measures, which will be used to assess the researcher models, from those measurements. The ratio of accurate predictions to the total amount of data sets is how accuracy is measured. Precision is the percentage of correctly identified positives that the authors correct. It is written as (1):

$$Precision = TP/TP + FP \tag{1}$$

Recall is used to gauge the accuracy of the classifier's predictions and is defined as (2):

$$Recall = TP/TP\ +\ FN \tag{2}$$

F-Score it is a measurement that considers both recall and precision, and it is calculated as (3):

$$F1 = 2 \times precision\ \times\ recall\ /\ precision\ +\ recall \tag{3}$$

Instead of using the conventional bag of words, the Afan Oromo news text dataset was preprocessed and embedded with the creation of a dense vector space for each word. The dataset was split into two sets: 20% for testing and 80% for training. Both the accuracy and data loss changes are pretty noticeable when running both cases for the number applied for each epoch and a batch size of 128.

5.1. NB classifier

Afan Oromo's fake news detection and classification, the researcher investigated various feature extraction techniques, including a bag of words, TF, and TF-IDF, and changed the size of the n-gram from n=1-4. Additionally, the authors changed the feature count, which ranged from 6,000 to 60,000. Table 3 displays the prediction outcomes for the various classification techniques. For classification using discrete characteristics like word counts for text classification, the multinomial NB classifier is appropriate. Typically, integer feature counts are required by the multinomial distribution. However, in actuality, fractional counts like TF-IDF might also be effective.

Table 3. Prediction outcomes for the various classification techniques

N-gram size	Term frequency			Term frequency inverse document frequency		
	6,000	10,000	60,000	6,000	10,000	60,000
Unigram	97.3	96.3	97.3	95.8	93.3	92.2
Bigram	94.2	94.0	94.0	92.9	88.9	75.2
Trigram	79.4	80.2	82.6	77.6	75.1	66.8
Four grams	72.4	73.2	76.9	71.9	69.4	65.6

To determine how similar the documents are, the term frequency method in this experiment counts words that appear in both datasets. The word counts are contained in an equal-length vector that represents each document. Each vector is then normalized so that the addition of its members will equal one. The likelihood that each word appears in the papers is then calculated from the word counts.

5.2. KNN classifier

Using a similarity metric, the KNN used to categorize texts determined the distance between each document and every other document in the training set. Finding the K closest neighbors among all training materials, the document is then assigned to the category with the most papers in the KNN set. It can be challenging to choose the value of k, but in practice, the general approach is to keep the value of k odd to prevent confusion between two classes of data. N refers to the number of samples in the training dataset [38]. The results of the KNN classifier are presented in Table 4.

5.3. Support vector machine classifier

The support vectors are the data points that are closest to the separating hyperplane. SVMs, also support-vector networks are the supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. For the given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane that categorizes new examples. The number of unique keywords is the dimension of the vector that the SVM approach uses to represent the text file. The results of the SVM classifier are presented in Table 5.

Table 4. Results of the KNN classifier

Parameter	Precision	Recall	F1-Score	Support
0.0	0.96	0.98	0.97	251
1.0	0.90	0.85	0.88	66
Average	0.95	0.95	0.95	

Table 5. Results of the SVM classifier

Parameter	Precision	Recall	F1-Score	Support
0.0	0.91	0.97	0.94	262
1.0	0.85	0.60	0.71	68
Average	0.92	0.92	0.90	330

5.4. Bidirectional long short-term memory

A BiLSTM, is a sequence processing model that consists of two LSTMs, one of which receives input forward and the other of which receives it reverse. The author employed Vanilla-RNN in the initial experiment, then LSTM in the second part of the experiment, and finally the BiLSTM technique in the final experiment. Table 6 presents the accuracy of each model.

Table 6. Accuracy of each model

Model	Accuracy result
Bidirectional-LSTM	0.91
Vanilla-RNN	0.90
LSTM	0.89

5.5. Comparing the results with the existing methods

In the existing and proposed methods, neural networks and SVM give two distinct results. The results of both procedures are compared to those of existing techniques. The parameters and time needed by the existing technique and the proposed hybrid technique are compared. The results of the neural network and SVM in the suggested hybrid technique are also compared, and the results with the accuracy rate are taken into consideration. The results of the comparison of the various classifiers are presented in Table 7.

Table 7. Comparison between the current/existing method and the proposed/new method

Parameter	Current/existing method		Propose/new method	
	SVM	NN	SVM	NN
Accuracy	0.88	0.92	0.92	0.95
Recall	0.90	0.91	0.92	0.95
F-Measure	0.89	0.89	0.90	0.95
Precision	0.91	0.90	0.92	0.95

6. DISCUSSION

Studies on fake news detection in various languages have been conducted; however, the findings are invalid for Afan Oromo because of its unique syllable structure, syntax, and semantics. Several knowledge-based, linguistic, and style-based false-news detection techniques have been developed. Fake news identification is a relatively recent field in Afan Oromo and other morphologically rich languages, and training and evaluating ML systems using easily accessible datasets is challenging. Despite Afan Oromo's widespread use in Ethiopia, insufficient research has been conducted to recognize bogus news on social media platforms, such as Facebook.

This study utilized the following measures, which are frequently used to assess classifiers in related fields, to assess the effectiveness of false news detection techniques: accuracy, precision, recall, and F1. The following conditions were used for all tests. NB and SVM produced better results, as explained and tested using the aforementioned data. The study achieved an accuracy of 94.1% using naive Bayesian and SVM. It is also obvious that unigram analyzers with feature values achieve improved accuracy. Additionally, the authors can see that employing all of the classifiers reduces the accuracy as the n-gram size increases. The author improved the naive Bayesian classifier's accuracy by extracting phrase frequency features at the unigram level. However, a document's wording and organizational structure are lost in the bag of words. The model simply considers the presence or absence of known words in both datasets.

The term frequency-inverse document frequency produced positive results in SVM, demonstrating the superiority of these systems over NB algorithms as classifiers. In a naive Bayesian classifier using all n-grams with all feature values, TF outperformed TF-IDF, whereas in a SVM classifier using all n-grams with all feature values, TF-IDF outperformed TF. The phrase "frequency" received a 62.5% unigram score in a SVM. With the word frequency feature, the left receives a score of 61.2% across all feature sizes. With all n-gram words except unigram and term frequency features on all feature values, the lowest accuracy of 42.3% was obtained. Figure 6 shows the various classification methods' prediction results.

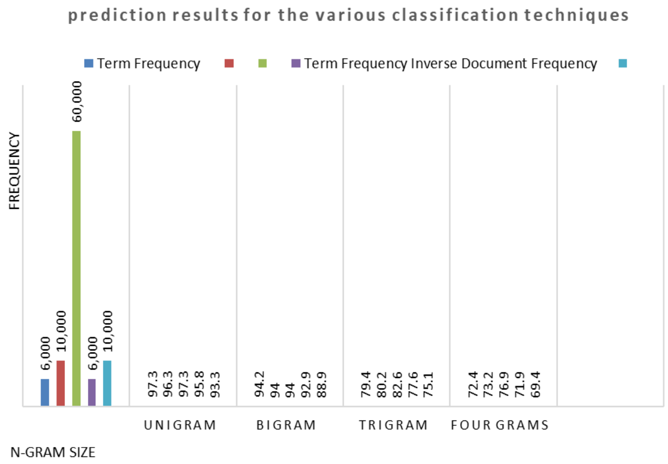

Figure 6. Various classification methods' prediction results

7. CONCLUSION

Fake news classification is a process that assigns predetermined categories to free-text content, but it faces challenges in identifying and categorizing false news articles in the Afan Oromo language. This research aims to develop an automated system that can efficiently and accurately identify and categorize deceptive information on social media platforms in the Afan Oromo language. The objective is to propose a methodology for identifying fake news on social media by utilizing Afan Oromo news text and deep learning techniques. The study proposed a model that uses ensemble approaches to differentiate between fake and authentic Afan Oromo language news articles. Deep learning algorithms require a substantial amount of data, but acquiring an adequate dataset poses a significant challenge in the Afan Oromo language. The inclusion of additional data in the news dataset enhances performance consistency and increases user confidence in the system.

The Bi-LSTM model is presented and proposed as a system prototype for future research using Afan Oromo news text datasets and other local language datasets in Ethiopia. The challenge with this strategy is that, in addition to multi-class prediction, the knowledge base would need to be continually and manually updated to be up to date. Some of the recommendations and planned work for additional research and improvement are as follows: the stop-word list employed in this study was developed during data processing and is largely news-related. The existence of a standard stop-word list would undoubtedly aid research in the domain of fake news prediction and other classification techniques; thus, a standard Afan Oromo stop-word list should be produced. The learning-based strategy was used in this investigation. Future work could be knowledge-based to increase trust in the system. Fake news detection based solely on supervised models on the text has proven insufficient in all circumstances. The majority of the research focuses on extra information to overcome this challenge. We believe that the most successful strategy would be to force the model with some kind of knowledge base, with the model's job being to extract information from the text and validate the information in the database. The issue with this strategy is that the knowledge base would have to be regularly and manually updated to stay current.

REFERENCES

[1] K. L. Arega, M. K. Alasadi, A. J. Yaseen, A. O. Salau, S. L. Braide, and J. O. Bandele, "Machine learning based detection of fake Facebook profiles in Afan Oromo language," *Mathematical Modelling of Engineering Problems*, vol. 10, no. 6, pp. 1987–1993, 2023, doi: 10.18280/mmep.100608.

[2] W. B. Demilie and A. O. Salau, "Detection of fake news and hate speech for Ethiopian languages: a systematic review of the approaches," *Journal of Big Data*, vol. 9, no. 1, pp. 1–17, 2022, doi: 10.1186/s40537-022-00619-x.

[3] P. Bugade, P. Sarode, and T. Pimple, "Fake news detection using machine learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 8, no. 4, pp. 4024–4027, 2021.

[4] R. Kunapareddy, S. Madala, and S. Sodagudi, "False content detection with deep learning techniques," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 5, pp. 1579–1584, 2019.

[5] R. Zafarani and X. Zhou, "Fake news detection," in *Companion Proceedings of the 2019 World Wide Web Conference,* 2019.

[6] N. Baarir and A. Djeffal, "Fake news detection using machine learning," in *IEEE 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*, 2021, pp. 125–130.

[7] D. K. Gurmessa, A. O. Salau, and A. Gedefa, "Afaan Oromo language fake news detection in social media using convolutional neural network and long short term memory," *Journal of Electrical and Electronics Engineering*, vol. 15, no. 2, pp. 37–42, 2022.

[8] P. Shah and Z. Kobti, "Multimodal fake news detection using a cultural algorithm with situational and normative knowledge," *2020 IEEE Congress on Evolutionary Computation (CEC),* Glasgow, UK, 2020, pp. 1-7, doi: 10.1109/CEC48606.2020.9185643.

[9] E. M. Okoro, B. A. Abara, A. O. Umagba, A. A. Ajonye, and Z. S. Isa, "A hybrid approach to fake news detection on social media," *Nigerian Journal of Technology*, vol. 37, no. 2, pp. 454–462, 2018, doi: 10.4314/njt.v37i2.22.

[10] R. A. Monteiro, R. L. S. Santos, T. A. S. Pardo, T. A. de Almeida, E. E. S. Ruiz, and O. A. Vale, "Contributions to the study of fake news in Portuguese: new corpus and automatic detection results," in *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018*, 2018, pp. 324–334, doi: 10.1007/978-3-319-99722-3_33.

[11] A. A. Alemu, M. D. Melese, and A. O. Salau, "Towards audio-based identification of Ethio-Semitic languages using recurrent neural network," *Scientific Reports*, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-46646-3.

[12] M. S. H. Ameur and H. Aliane, "Arabic COVID-19 multi-label fake news and hate speech detection dataset," *Procedia Computer Science*, vol. 189, pp. 232–241, 2021, doi: 10.1016/j.procs.2021.05.086.

[13] M. C. Buzea, S. Trausan-Matu, and T. Rebedea, "Automatic fake news detection for Romanian online news," *Information (Switzerland)*, vol. 13, no. 3, pp. 1–13, 2022, doi: 10.3390/info13030151.

[14] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: a data mining perspective," *International Journal of Innovative Technology and Exploring Engineering*, 2020.

[15] Gazette Federal Negarit, Federal Negarit Gazette of the Federal Democratic Republic of Ethiopia, 1998.

[16] M. A. Chekol, M. A. Moges, and B. A. Nigatu, "Social media hate speech in the walk of Ethiopian political reform: analysis of hate speech prevalence, severity, and natures," *Information Communication and Society*, vol. 26, no. 1, pp. 218–237, 2023, doi: 10.1080/1369118X.2021.1942955.

[17] T. G. Debela, "Afan Oromo search engine: a thesis submitted to the school of graduate studies of the Addis," M.S. thesis, Department of Computer Science, Addis Ababa University, Ethiopia, 2010.

[18] E. F. Dinsa and R. B. P, "Application of data mining classification algorithms for Afaan Oromo media text news categorization," *International Journal of Computer Trends & Technology*, vol. 67, no. 7, pp. 73–79, 2019, doi: 10.14445/22312803/ijctt-v67i7p112.

[19] T. Debela, "Designing a stemmer for Afaan Oromo text: a hybrid approach," M.S. thesis, Addis Ababa University, Ethiopia, 2010.

[20] A. A. Alemu, M. D. Melese, and A. O. Salau, "Ethio-semitic language identification using convolutional neural networks with data augmentation," *Multimedia Tools and Applications*, vol. 83, no. 12, pp. 34499–34514, 2024, doi: 10.1007/s11042-023-17094-y.

[21] T. F. Berhanu, "Afaan Oromo automatic news text summarizer based on sentence selection function," M.S. thesis, Addis Ababa University, Ethiopia, 2013.

[22] T. Granskogen, "Automatic detection of fake news in social media: using contextual information," M.S. thesis, Department of Computer Science, Norwegian University of Science, Trondheim, Norway, 2018.

[23] D. Gurmessa, "Afan Oromo text content-based fake news detection using multinomial naive Bayes," *International Journal of Innovations in Management Science, and Engineering (IJIMSE)*, pp. 26–37, 2020.

[24] J. Shaikh and R. Patil, "Fake news detection using machine learning," *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, Gunupur Odisha, 2020, pp. 1-5, doi: 10.1109/iSSSC50941.2020.9358890.

[25] R. R. Mandical, N. Mamatha, N. Shivakumar, R. Monica and A. N. Krishna, "Identification of fake news using machine learning," *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, Bangalore, India, 2020, pp. 1-6, doi: 10.1109/CONECCT50063.2020.9198610.

[26] A. A. Baykara, "Fake news detection using machine learning and deep learning algorism," in *2020 International Conference on Advanced Science and Engineering (ICOASE),* Duhok, Iraq, 2020, pp. 18-23, doi: 10.1109/ICOASE51841.2020.9436605.

[27] S. Aphiwongsophon and P. Chongstitvatana, "Detecting fake news with machine learning method," in *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Chiang Rai, Thailand, 2018, pp. 528-531, doi: 10.1109/ECTICon.2018.8620051.

[28] G. De Magistris, S. Russo, P. Roma, J. T. Starczewski, and C. Napoli, "An explainable fake news detector based on named entity recognition and stance classification applied to," *Information MDPI*, vol. 13, no. 3, pp. 1–14, Mar. 2022, doi: 10.3390/info13030137.

[29] 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 System in Distributed and Cloud Environments*, 2017, pp. 127–138, doi: 10.1007/978-3-319-69155-8_9.

[30] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, Kyiv, UKraine, 2017, pp. 900-903, doi: 10.1109/UKRCON.2017.8100379.

[31] S. Bajaj, "The pope has a new baby fake news detection using deep learning," Stanford University, 2017.
[32] A. Thota, "Fake news detection: a deep learning approach," *SMU data science review*, vol. 1, no. 3, pp. 1–21, 2018.
[33] G.O. Ganfure, "Comparative analysis of deep learning based Afaan Oromo hate speech detection," *Journal of Big Data*, vol. 9, 2022, doi: 10.1186/s40537-022-00628-w.
[34] B. Al-Ahmad, A. M. Al-Zoubi, R. A. Khurma, and I. Aljarah, "An evolutionary fake news detection method for COVID-19 pandemic information," *Symmetry*, vol. 13, no. 6, pp. 1–6, 2021, doi: 10.3390/sym13061091.
[35] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," *WWW '11: Proceedings of the 20th international conference on World wide web*, 2011, 675-684, doi: 10.1145/1963405.196350.
[36] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: a data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.
[37] L. Buitinck *et al.*, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[38] H. F. Villela, F. Corrêa, J. S. de A. N. Ribeiro, A. Rabelo, and D. B. F. Carvalho, "Fake news detection: a systematic literature review of machine learning algorithms and datasets," *Journal on Interactive Systems*, vol. 14, no. 1, pp. 47–58, 2023, doi: 10.5753/jis.2023.3020.

BIOGRAPHIES OF AUTHORS

Dr. Ayodeji Olalekan Salau received the B.Eng. in electrical/computer engineering from the Federal University of Technology, Minna, Nigeria. He received the M.Sc. and Ph.D. degrees from the Obafemi Awolowo University, Ile-Ife, Nigeria. His research interests include research in the fields of computer vision, image processing, signal processing, machine learning, control systems engineering and power systems technology. He serves as a reviewer for several reputable international journals. His research has been published in many reputable international conferences, books, and major international journals. He is a registered engineer with the Council for the Regulation of Engineering in Nigeria (COREN), a member of the International Association of Engineers (IAENG), and a recipient of the Quarterly Franklin Membership with ID number CR32878 given by the Editorial Board of London Journals Press in 2020 for top quality research output. More recently, his research paper was awarded the best paper of the year 2019 in Cogent Engineering. In addition, he is the recipient of the International Research Award on New Science Inventions (NESIN) under the category of "Best Researcher Award" given by Science Father with ID number 9249, 2020. Currently, his works at Afe Babalola University in the Department of Electrical/Electronics and Computer Engineering. He can be contacted at email: ayodejisalau98@gmail.com.

Mr. Kedir Lemma Arega is currently a Ph.D. student in software engineering at Addis Ababa Science and Technology University. He works as an ICT infrastructure and administrator team leader at Ambo University, Woliso Campus, Ethiopia. His research interests include research in the fields of computer vision, image processing, signal processing, machine learning, and natural language processing. He can be contacted at email: kedirnaw1999@gmail.com.

Dr. Ting Tin Tin received her B.Sc. and Ph.D. in computer sciences from University of Science, Malaysia. She joined Gemalto as telecommunication software engineer in Singapore before joined academic industry after her Ph.D. graduation. She has more than 12 years of lecturing, supervising projects, and research. She received her professional certification in project management from PMI and data analytics from SAS. Currently, she is attached to INTI International University responsible primary in research and postgraduate supervision. She is a passionate researcher in her publication collaborate with national and international researchers i.e., Singapore, Bangladesh, India, Nigeria, Pakistan, United Arab Emirates, and Saudi Arabia. Her research interests including big data analytics, information systems engineering, educational data mining, psycho-academic research, and software engineering. She started being active in her research since 2022, increasing her google scholar H-index from 1 to 4 (2023). Currently, she is a freelancer serves as freelance lecturer in INTI International University, Monash University, Tunku Abdul Rahman University of Management and Technology, and Methodist College Kuala Lumpur. Her responsibility including Ph.D. students' supervision; master's programme course leader; course leader; delivery, coordination and development of learning materials for undergraduates, setting of examination papers and marking schemes, conduct lecture, tutorial, practical laboratory, mid-term test; open day faculty consultant; final year project supervisor and moderator; research publication. She can be contacted at email: tintin.ting@newinti.edu.my.

Mr. Andrew Quansah is a Ph.D. student in the Department of Electrical and Computer Engineering at the University of North Carolina at Charlotte. He holds an M.Sc. microelectronics and wireless intelligent system from Coventry University in the UK. He received his BSc. in electrical and electronics engineering from the Kwame Nkrumah University of Science and Technology. His research interest includes advanced embedded systems and computer architecture, wireless communication systems, real time and application artificial intelligence, real time, and embedded operating systems. He can be contacted at email: aquansa1@uncc.edu.

Mr. Kwame Sefa-Boateng is currently pursuing a master's in information technology with a concentration in advanced data and knowledge discovery at the University of North Carolina at Charlotte. He holds a master's in information and communication technology from the School of Technology, Ghana Institute of Management and Public Administration in Ghana. He received his bachelor's in computer science from the University of Ghana. He's interested in the researching automation of service-level agreements in cloud computing and data analytics. He can be contacted on email: ksefaboa@charlotte.edu.

Ismatul Jannat Chowdhury is currently immersed in the pursuit of a Ph.D. in electrical engineering at the esteemed University of North Carolina at Charlotte, USA. Her academic journey encompasses a master's degree in data science from the University of The West of England (UWE) in Bristol, UK, and a Bachelor's degree in electrical and electronics engineering from Premier University, Chittagong, Bangladesh. Her diverse research interests span across machine learning and artificial intelligence, data science, optics and photonics, wireless networking, sensing, and security. This broad spectrum reflects her dedication to staying at the forefront of technological advancements. Currently engaged in a Ph.D. program, Ismatul channels her commitment into advanced research and analysis in these cutting-edge fields. She can be contacted at email: ichowdhu@uncc.edu.

Dr. Sepiribo Lucky Braide is an Associate Professor in the Department of Electrical Electronics Engineering, Rivers State University, Port Harcourt, Nigeria. He is the past Head of Department, Electrical Engineering from 2018 to 2021, fellow of the Nigeria Institute of Electrical Electronics Engineers (FNIEEE) and currently the Post Graduate Coordinator in the Department, from 2021 till date. He is a member of several professional bodies/organizations among many includes; Institute of Electrical Electronics Engineers (IEEE), Nigeria Society of Engineers (MNSE), Council for the Regulation of Engineering in Nigeria (COREN), Nigeria Institute of Electrical Electronics Engineers (MNIEEE), International Association of Engineers (MIAENG), he had the following degrees in Electrical Engineering profession which includes; bachelor of technology (B.Tech.), master of technology (M.Tech.), and doctor of philosophy (Ph.D.). He has attended several conferences these includes: ICEPT, NSE, COREN, IEEE, WCECS (San Francisco, USA Oct., 2019), WCE (London UK, July, 2019). His strong pragmatic confrontation to analyses of finding solutions to challenging engineering problems made him compete in many highly reputable international journals, most recently in 2018, International Journal Engineering and Science Invention (IJESI) classified one of his article as Best Paper Award - 2018 certified and titled: "A Mathematical Model of Double Exponential Wave Shape (Impulse Generator) for Power Sub-station using Laplace Transform". He has received so many honors and awards both local and international. He can be contacted at email: braidesepiribo@yahoo.com.