SocialGuard: Bangla Text-Based Gender Identification for Enhancing Integrity in Social Networks

Md Jahangir Alam*, Sultan Ahmed†, Ismail Hossain*, Sai Puppala*, Zahidur Talukder‡, Sajedul Talukder*
*School of Computing, Southern Illinois University, Carbondale, IL, USA
†University of Maryland, Baltimore County, USA
‡The University of Texas at Arlington, USA
sajedul.talukder@siu.edu

2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC) | 979-8-3503-7696-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/COMPSAC61105.2024.00086

***Abstract*—In this study, we address the task of discerning gender through the textual content of social media, a crucial step in detecting and mitigating counterfeit account activity. Ensuring accurate gender portrayal on digital platforms is essential for creating a secure and inclusive cyberspace. While research exists for languages like English, Russian, and Arabic, Bangla remains underexplored. To address this, we compiled 15,000 Bangla posts from Facebook groups, profiles, pages, blogs, and forums. We trained seven traditional machine learning algorithms (NB, SVM, LR, DT, RF, SGD, KNN) and three deep learning models (MLP, LSTM, GRU), using stylometric features, Term Frequency-Inverse Document Frequency (TF-IDF), and word embeddings. Traditional models generally outperformed deep learning models, except with stylometric features. Notably, the Stochastic Gradient Descent (SGD) model with TF-IDF achieved the highest accuracy (78.33%) and F1-Score (87.67%). Additionally, Continuous Bag of Words (CBOW) outperformed Skip-Gram (SG) in training the word2vec model, with top accuracy and F1-Score of 75.13% and 79.92%, respectively. These findings represents a significant stride forward in the field of gender identification from Bangla text.**

***Index Terms*—Social network, identity attack, gender spam, supervised detection, Bangla language**

I. Introduction

The rise of fake and anonymous accounts on social media and other online platforms is a growing issue, with many profiles being unverified. This trend affects more than just identity misrepresentation. Most online content, from social media posts to news articles, blogs, and emails, is text-based. This situation highlights the importance of Author Profiling (AP), a method that uses shared information to infer an author's demographic and psychological traits, such as gender, age, personality, native language, or political leanings [1]. AP's applications are varied and significant, helping in identifying misinformation, determining who wrote anonymous texts, recognizing style differences for targeted ads, spotting unusual behavior, and exposing state-backed troll accounts [2], [3], [4]. AP's effectiveness mainly comes from analyzing text and the author's unique writing style [5], playing a crucial role in creating a more secure and genuine online environment.

Among the world's languages, Bangla stands out due to its extensive native speaker base. As the fifth most spoken native language and a significant Indo-European language, Bangla has approximately 300 million native speakers and an additional 37 million second-language speakers [6], [7]. It serves as the lingua franca for 98% of the population in Bangladesh and is recognized as the official national language. The influence of Bangla extends beyond the geographical boundaries of Bangladesh, with significant diasporas in the Middle East, Europe, and the United States [7]. The digital revolution, particularly the initiative of Digital Bangladesh [8], has further amplified the presence of Bangla in the online realm. Platforms like Facebook, LinkedIn, and Twitter see substantial use of Bangla, with Facebook alone boasting 33.71 million active users in Bangladesh [9]. This surge in digital communication in Bangla necessitates the development of robust NLP tools tailored to this language.

Despite the critical need and the vast speaker base, research in gender identification from Bangla texts remains surprisingly sparse, particularly when contrasted with efforts in languages such as English, Russian, and Arabic. To address this gap, we extend our earlier work [10] in which we experiment on 5,000 posts utilizing several traditional machine learning models for gender identification. In this study we perform rigorous experimentation of stylometric analysis, TF-IDF, and word embedding techniques for gender identification in Bangla texts. Utilizing a dataset of 15,000 posts collected from various online Bangla sources, we embark on a comparative analysis of seven traditional machine learning models (NB, SVM, LR, DT, RF, SGD, KNN) and three deep learning-based models (MLP, LSTM, GRU). Our findings are significant: they suggest that while deep learning models hold promise, traditional machine learning models, particularly the Stochastic Gradient Descent (SGD), demonstrate superior performance in certain contexts, such as with TF-IDF features. Finally we design a system named SocialGuard (see Figure 1), develop a social network prototype and simulate adversary

2836-3795/24/$31.00 ©2024 IEEE
DOI 10.1109/COMPSAC61105.2024.00086

model i.e creating posts, validating user gender.

Our work not only contributes to filling a crucial research void in Bangla NLP but also sets a precedent for future explorations in this area. By establishing a framework for gender identification from Bangla text, we pave the way for more nuanced and culturally aware computational models, thereby enhancing the relevance and effectiveness of AP techniques in diverse linguistic landscapes. The main contributions of the paper are as follows:

- We created a dataset that contains user text along with the user's gender information in the Bangla language.
- We designed a system that aims to address the identity deception attacks originating from the "Fake Gender Social Media Account" by identifying the author gender from the Bangla text.
- We propose several types of stylometric features as gender indicators for the Bangla language. We designed a set of measures to infer the gender of the author from short writings through extensive experiments.
- We demonstrate superior accuracy of the CBOW model as compared to the Skip Gram model for gender extraction of users from the Bangla dataset. We also exhibit superior accuracy of TF-IDF features as compared to Word2Vec and Stylometric features.
- Finally, doing a comparison with different ML models, we show traditional machine learning algorithms are performing better than deep learning algorithms.

The rest of the paper is organized as follows. Section II introduces the key Research Questions that underpin our study and the related works. Section III presents the system and adversary model, establishing the foundational framework. The methodology of of the proposed solution are detailed in Section IV. Section V delves into the experiment details, followed by Section VI, which demonstrates our proposed model's robustness. Section VII describes the ethical considerations we kept in mind while performing the data collection and experiment. The discussions and limitations of our study are examined in Section VIII. Finally, Section IX concludes the paper, summarizing our key findings and suggesting avenues for future research.

II. Related Works

Initial investigations by Corney and Anderson [11] in gender identification from email data used Support Vector Machines (SVM), achieving up to 71.2% accuracy. Subsequent studies by Fink et al. [12] and others [13] expanded to social media and blogs, exploring various profiling features.

Research by Sboev et al. [14] employed LSTM and GRU networks with morphological and syntactic features for gender and sentiment identification in Russian texts. In Arabic contexts, studies by Mubarak et al. [15] and Shalabi et al. [16] focused on gender identification from online articles and Twitter using bag-of-words and stylometric features.

Bsir et al. [17] used a Gated Recurrent Unit Deep learning model with lexical features for gender detection in Facebook and PAN corpus, achieving 62.1% and 79% accuracy. Cheng et al. [18] focused on English datasets, applying stylometric features with traditional machine learning algorithms, reaching 73% to 83% accuracy.

Rahman et al. [19] constructed a Bengali speech corpus for gender identification. Tripto et al. [20] analyzed YouTube Bangla comments for sentiment identification using LSTM and CNN, with a maximum accuracy of 65.97%. Nia et al. [21] improved gender identification accuracy by combining the ViT model for images and the Bert model for text.

To the best of our knowledge, our work is the first to focus specifically on gender identification from Bangla social media texts using TF-IDF feature representation and word embeddings, and has outperformed previous Bangla studies in gender identification.

Research Objective.

Throughout this paper, we address these research questions (RQs), each aiming to delve deeper into the complexities and nuances of gender identification in Bangla texts, offering insights into both the technical and cultural dimensions of this research:

- **RQ1:** How can accurate gender identification enhance safety & integrity on social media platforms, particularly for Bangla language speakers?
- **RQ2:** Can we design a system that can accurately predict the gender of authors based on their social media text content?
- **RQ3:** What are the unique challenges in processing Bangla texts for gender identification compared to other languages?
- **RQ4:** What are the comparative performances of traditional machine learning models and deep learning models in the task of author gender classification in Bangla texts?
- **RQ5:** How does TF-IDF approach compare to word embeddings in identifying gender-specific writing styles in Bangla social media texts?

Each question aims to delve deeper into the complexities and nuances of gender identification in Bangla texts, offering insights into both the technical and cultural dimensions of this research.

III. System and Adversary Model

We consider a Facebook-like online social network that primarily operates by allowing users to generate and consume content. One of the most common content types is a "post." The propagation system of these posts hinges on a combination of user-generated content, engagement metrics, and algorithmic determinations. When a user shares a post, its initial visibility is to their most frequent connections. The algorithm gauges the post's popularity

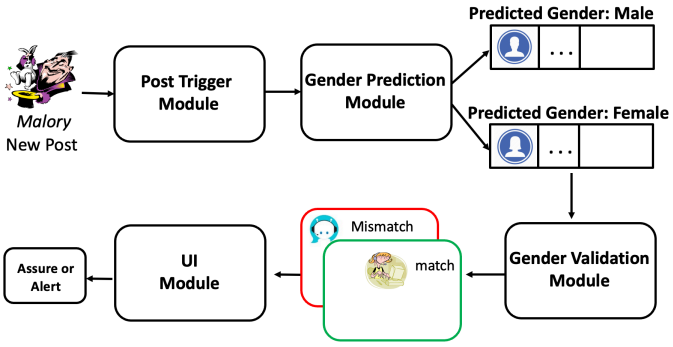

Fig. 1: SocialGuard System architecture. Post Trigger Module captures text posts, the Gender Prediction Module analyzes them, the Gender Validation Module compares predictions with declared gender, and the UI Module displays relevant labels.

through user interactions such as likes, comments, and shares. Posts that garner high engagement are presented to a wider audience, possibly even beyond the user's direct connections. Recommendations can further amplify a post's reach, especially if its content aligns with the interests of a broader user base. The system's spread mechanics can be influenced by factors like user activity, content relevance, and ad promotions.

System Model. In addressing RQ2, we have designed SocialGuard, a system structured around four integral modules. Each module plays a pivotal role in identifying, predicting, validating, and alerting the gender of a post's author (see Figure 1). This layered approach not only increases the accuracy of gender prediction but also offers a robust defense against identity deception. Let's delve into the specifics of each module:

- **Post Trigger Module**: Activated upon a user's creation of a new textual post, this module immediately captures and processes the content for further analysis.
- **Gender Prediction Module**: Leveraging a refined gender classifier, this module receives the post's content and predicts the likely gender of the author, either male or female. We discuss this module in detail in the later sections.
- **Gender Validation Module**: Serving as a validation checkpoint, this module cross-references the predicted gender against the gender stated in the user's profile which answers the RQ1. For this, SocialGuard gets the user gender information by scraping the user profile page. Validation module generates alert to be shown alongside the post if it finds mismatch between scraped and predicted gender as shown in Figure 2.
- **UI Module**: As the final step, this module generates pertinent labels based on prior analyses. These labels, displayed alongside the post, either offer validation to the user or issue cautionary alerts, depending on the congruence of the predictions and the stated gender. Figure 2 illustrates the envisioned operation of the Gender Validation Module coupled with the UI Module, which together analyze and flag potential misrepresentations of gender identity on social media. In this figure we if the username represents a Male username where the content actually represents a Female author. Normally the reader of this post will assume that the post is authored by the Male user who is the owner of the Facebook account from which the post has been posted. In this scenario, our system will detect the actual gender from the text and will show the predicted gender on the UI.

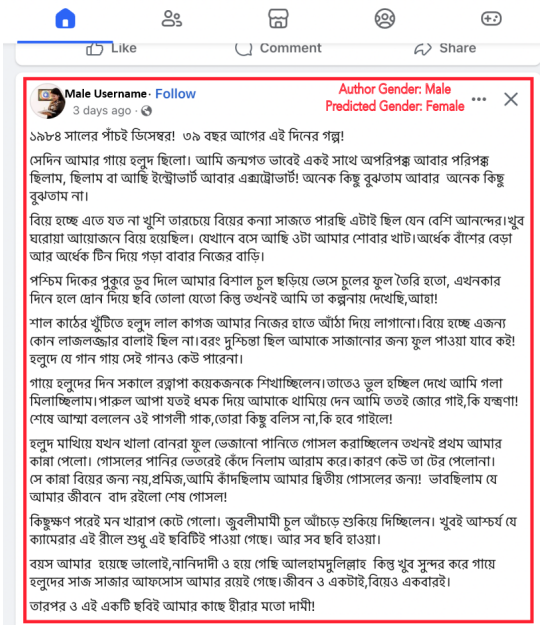

Fig. 2: Demonstration of Gender Validation and UI Modules in action, processing & presenting gender predictions in Bangla posts to caution against identity misrepresentation.

Adversary Model. We simulate an adversary model by creating fake social media accounts with specified genders and posting content as adversaries. This is followed by performing gender prediction and validation on these accounts. We consider Social Media users as adversaries who create and manage "Fake Gender Social Media Accounts", a.k.a deceptive profiles, created using fraudulent information that falsely represents their gender. The portrayal of a fake gender is primarily a manipulation tool, designed to exploit biases and build undue trust. Adversaries use fake accounts to create posts that can reach victim accounts. While the adversary operates, it's assumed that the platform's security doesn't immediately detect the account, and the targets—ordinary social media users—aren't necessarily well-versed in recognizing such threats. The spectrum of attacks they can launch is broad. These range from "catfishing" [22]—where individuals are lured into emotional or financial traps—to sophisticated influence operations aiming to sway public opinion. Fur-

thermore, these adversaries may resort to cyberbullying, engage in reputation sabotage through false testimonies, or even propagate malware [23], [24] and phishing campaigns [25]. Affiliate fraud and gender-targeted scams [26] also become feasible, given the trust that the fake gender might elicit from unsuspecting victims.

IV. Methodology

In this section we discuss the methodologies of word vector formation from user texts for the gender prediction. To convert each text to a particular input format, we examine three ways for feature extraction: the stylometric approach, the TF-IDF vectorizer approach, and the word embedding approach. Finally, we present the architecture of our model.

A. Word Vector Formation

We use a deep learning model in our proposed solution. To use the deep learning model, we convert our text to a word vector using stylometric features and word embedding representation approaches.

1) ***Stylometric Features Approach****:* Stylometric features are the features that capture the writing style of different authors of both genders. We compute a large set of stylometric features based on existing works of [11]. These features are categorized into four types: lexical features, structural features, syntactic features, and content-specific features.

Lexical features. These are the most common set of stylometric features that are intended for stylistics and text readability analysis. These features also signify language assessment and first and second language acquisition. Lexical features consist of character-based and word-based features. These features are concerned with the usage frequency of individual letters, vocabulary richness, entropy measure, the consecutive occurrence of words, etc.

Syntactic features. These are primarily intended for identifying writing formation patterns such as the usage of punctuation marks. These features include the total number of commas, colons, question marks, exclamation marks, etc. Syntactic features are useful in deriving gender from text because of men's and women's different habits of using punctuation. For example, women tend to use more question marks than men [27].

Structure-based features. They focus on the way of organization of the layout of a text by an author. The organization of articles represents different habitual facts of an author such as paragraph length and use of greetings. As online texts have less content information but richer stylistic information, these habits are seen to be more prominent in these texts in bearing strong authorial evidence of personal writing styles. We compute 8 structure-related features.

Content-specific features. They represent domain-specific terms. For these features, we first collect the feature words as suggested for the Arabic language in [16]. Then we prepare the Bangla feature words by translating these Arabic words using Google Translator service API. We have translated Arabic words into 5 categories: Economy, Policy, Social, Sport, and Negative. This translation resulted in many duplicates, flaws, and inconsistencies in the translated lexicons. We clear all of these issues by manually inspecting the lexicons.

For each text of the user, the feature extractor produces a feature vector of a dimension of 141, which represents the values of the 141 stylometric features. As these feature sets contain information on the writing style of a user measured by various methods, the feature values can range from 0 to any positive value. As we want to ensure all features are treated equally in the classification process, we normalize the feature values using the min-max normalization method to ensure all feature values are between 0 and 1. We normalize the feature values using the equation below:

$$x'_{ij} = \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)} \tag{1}$$

where x_{ij} is the jth feature in the ith example, min (x_j) and max (x_j) are the minimum and maximum feature values of the jth feature respectively.

2) ***TF-IDF Vectorizer Approach****:* TF-IDF (term frequency-inverse document frequency) is a text vectorizer that converts the text of a user into a feature vector. After tokenizing a text we get a list of tokens or words. Each token is referred to as a term. In any document, the term frequency represents the number of occurrences of a term. On the other hand, document frequency represents the number of documents containing that term. Term frequency indicates the importance of a specific term in a document. Document frequency indicates how common the term is [28]. We implement TF-IDF vectorizer from `scikit-learn python` library. We take the most frequent 1000 words (tokens) to limit the number of features to be extracted from each document. As part of pre-processing, we remove letters other than the Bangla alphabet. For example, let us consider we have some texts, which have to be converted to feature vectors using a TF-IDF vectorizer. For converting these texts into feature vectors, we first identify unique words and count how many times these words occur in each text. Then we compute inverse document frequency (IDF) using the following formula:

$$idf_i = log\frac{n}{df_i} \tag{2}$$

where df_i represents how many documents contain the term i and n is the total number of documents. We calculate the inverse document frequency for each word. Then TF matrix is multiplied by the IDF score to get a vectorized form of each text. We convert all texts into vectors. These vectors can be fed into any machine-learning algorithm.

*3) **Word Embedding Representation Approach**:* Traditionally, the bag-of-words (BOW) model is used to transform the text into feature vectors in text classification [29]. In this model, a text is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity. Motivated by the recent success of deep learning models in text classification, we use word embedding as features instead of applying the BOW model. In this approach, individual words are represented as real-valued vectors in a predefined vector space. In the following, we explain 1) how the CBOW & Skip-Gram model of word2vec work, and 2) how the word2vec model is used to generate word embeddings from our dataset. Finally, we explain how our dataset is classified using these word embeddings.

word2vec. This is an efficient algorithm proposed by Google [30], which can learn a standalone word embedding from a text corpus efficiently maintaining the contextual meaning of words. word2vec has two model architectures: Continuous Bag of Words (CBOW) and Skip-Gram (SG). CBOW Model takes the context of each word as the input and predicts a target word corresponding to the context. SG predicts the surrounding window of context words based on the current single word. The word vector prediction is not influenced by the order of context words.

Word Embedding using word2vec. word2vec model can capture a lot of information maintaining semantic, conceptual, and contextual relations [31]. We learn the embedding vector of each word from user posts on Facebook using these CBOW and Skip-Gram models. For example, let us consider we have two sentences in our dataset: [I have a book (আমার একটি বই আছে), I have a car (আমার একটা গাড়ি আছে)]. We first split the sentences and generate a two-dimensional vector: [[I(আমার), have(আছে), a(একটি), book(বই)], [I(আমার), have(আছে), a(একটা), car(গাড়ি)]]. Then we pass this two-dimensional vector to the word2vec model. Skip-Gram and CBOW model generate the word vector from this two-dimensional dataset using a window size of 5. The size of the word vector is 300. We use the Gensim Python library to implement the word2vec model. This model returns a 300-dimensional vector for each of these words: I, have, a, book, car. We save these word vectors and later use these vectors as the weight of the embedding layer.

Classification. For classification, we first encode each word of the sentence S using a unique number. Then we multiply this encoded vector with the embeddings of words present in S to form the hidden representation of S. These sentence representations are used to train a classifier. Specifically, we use the softmax function to compute the probability distribution over the classes.

V. Experiment

A. Dataset

Data Collection. In our gender identification project focusing on Bangla social media texts, we adopted a comprehensive data collection approach, prioritizing ethical and legal considerations. We targeted popular Facebook groups, profiles, and pages with 1,000 to 5,000 members or followers. Using Facebook's Graph API, we gathered posts and comments while adhering to data usage policies and respecting user privacy. When direct API access was not feasible, we used automated techniques with Selenium and manual collection methods, ensuring compliance with privacy guidelines.

For forums and discussion boards, we used web scraping tools like Beautiful Soup and Scrapy to gather data from high-engagement threads popular among Bangla-speaking communities. For blogs, we leveraged platforms like WordPress, Blogger, Medium, and Google Blog Search to find relevant Bangla content, collecting data from RSS feeds and comments sections to capture conversational and opinion-rich texts.

Overall, we collected data from approximately 500 unique users, ensuring a comprehensive and diverse representation of Bangla language users. This broad-based approach helped us capture the varied nuances and expressions inherent in the Bangla language, making our dataset extensive and reflective of the diverse user base.

Throughout our data collection process, we were committed to ensuring the anonymity of personal data and respecting user privacy. We adhered to data protection laws, including GDPR, and local regulations in Bangladesh. Our scraping methods are designed to be non-intrusive and respectful of the target websites' server resources. In line with GDPR and local regulations in Bangladesh, we took

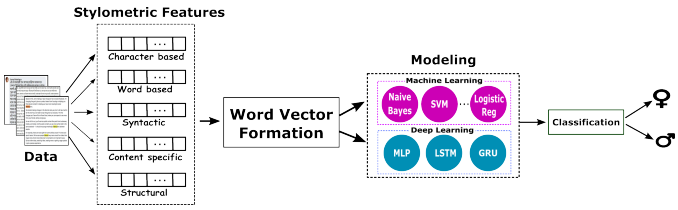

Fig. 3: Model Architecture for Stylometric Feature.

Fig. 4: Model Architecture for TF-IDF.

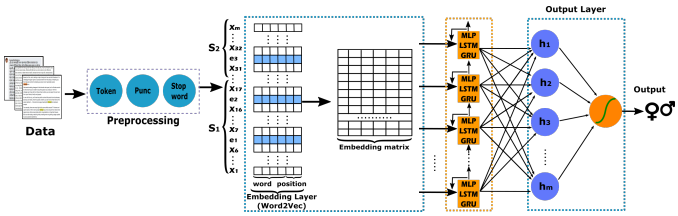

Fig. 5: Model Architecture for Word Embedding.

significant measures to de-identify personal information, including the application of cryptographic hash functions to anonymize Facebook account names. In total, we collected 15,000 posts, with 9,200 attributed to male authors and 5,800 to female authors. The dataset and their sources can be found at the following GitHub repository[1].

Multifaceted Data Authentication and Labeling Strategies. We implemented a robust and multifaceted approach to ensure the authenticity and relevance of our dataset. Initially, we conducted a rigorous filtering process, followed by a meticulous manual labeling procedure to accurately assign gender categories to the members. This labeling was grounded in an in-depth analysis of each member's Facebook activities, particularly focusing on their engagement levels and contribution points within their groups or platforms.

Enhanced Dual Authenticity and Relevance Verification. To ensure the legitimacy, relevance, and accuracy of our gender-based analysis data, we implemented a comprehensive two-pronged approach. First, we meticulously screened users' names for authenticity and evaluated the full-text content of their posts, applying strict inclusion and exclusion criteria to enhance dataset reliability. Additionally, we strategically collected data from well-known Bangla-speaking individuals with publicly confirmed genders. This approach minimized ambiguity in gender identification and leveraged publicly available information to verify gender, thereby reinforcing the integrity and authenticity of our gender-specific text analysis.

Comprehensive Originality Verification. We addressed the issue of duplicate and non-original text entries by implementing a filtering process to identify and remove such instances. We analyzed our dataset for internal duplicates and extended this verification to external sources by searching for identical text entries on Google and various social media platforms. This step was essential to ensure the texts, especially from well-known individuals with publicly confirmed genders, were original. We used a combination of manual searches and automated tools to scan for duplicates. When a match was found, we reviewed it to determine its originality. This thorough vetting process helped us eliminate replicated texts, enhancing the reliability and validity of our gender-specific text analysis.

Gender-Specific Keyword Filtering. To ensure the originality and authenticity of our dataset, we employed a targeted filtering approach using gender-specific keywords in Bangla. These keywords, selected based on linguistic research and cultural context, helped refine our dataset to enhance the accuracy of gender classification. By focusing on words and expressions predominantly used by one gender, we obtained valuable cues for our analysis. We remained cautious to avoid reinforcing stereotypes or biases by balancing and informing our keyword selection with a nuanced understanding of gendered language use.

[1]https://github.com/supreme-lab/gender-identification

In our data labeling process, multiple labelers were involved to mitigate biases and errors. Discrepancies in labeling were resolved through a thorough review and reconciliation process, significantly enhancing the accuracy and reliability of our dataset. This collaborative approach ensured a more robust and credible dataset for our analysis.

Pre-processing. In addressing RQ3, we tackled unique challenges specific to the Bangla language in social media, especially within Facebook group discussions. The informal, multifaceted nature of these posts includes various non-textual elements like URLs, images, tags, and links. Our pre-processing step involved refining these texts by meticulously removing non-Bangla alphanumeric characters, punctuation, URLs, images, links, hashtags, and user tags. This cleansing process, similar to our previous works [32], [33], ensured the text was suitable for analytical purposes, focusing on gender identification.

Another critical aspect of our pre-processing was the treatment of stop words. Stop words in any language typically comprise the most common words, which, although frequently used, often add little to no semantic value to the text. They can also introduce unnecessary repetition in the dataset. To tackle this, we tokenized the texts and systematically removed these stop words. Our reference for the Bangla stop words was a comprehensive list available in a GitHub repository [34], as mentioned in the research work of Tripto and Ali [20]. It is important to note that while Bangla stop words are not universally known, the specific ones utilized in our study are detailed in the aforementioned repository, providing transparency and reproducibility in our research methodology.

An interesting linguistic feature we encountered in our analysis was the elongation of words, a form of expression often used to convey stronger emotions or emphasis [35]. For example, the phrase খুউউউব সুন্দর" (very beautiful) with its elongated vowels carries a deeper emotional resonance compared to its standard form খুব সুন্দর". Our observations revealed that such elongations, particularly common among female users, are significant in conveying sentiments and, consequently, in the context of gender identification. Acknowledging the value of these linguistic nuances, we consciously decided against applying lemmatization to the texts. This approach ensured that we preserved the emotional and contextual richness inherent in word elongations, which could be pivotal in distinguishing gender nuances in Bangla texts.

Through these tailored pre-processing steps, our objective was to distill the dataset to its most analytically valuable form, ensuring that it accurately mirrored the authentic linguistic patterns relevant to our study on gender identification from social media texts.

B. Model Architecture

We implement two different approaches to gender identification. The first one uses deep learning model archi-

tecture and the one is traditional model architecture. In the deep learning model architecture, we implement Multi-Layer Perceptron (MLP), Long Short Term Memory (LSTM) model, and Gated Recurrent Unit (GRU) models. In the traditional model architecture, we implement Support Vector Machine, Naive Bayes, K-Nearest Neighbors, Random Forest, Decision Tree, Stochastic Gradient Descent, and Logistic Regression models. Figures 3, 4, 5 show the architecture of our proposed model for three different features. Traditional models are implemented to compare the performance of the deep learning models.

Models with Stylometric Features. We prepare word vectors using the stylometric feature approach discussed in Section IV-A1. These vectors are passed to the LSTM layer having 300 nodes. The output of the LSTM layer is passed to a dense layer. We use sigmoid [36] as an activation function. The optimizer is RMSprop and binary cross entropy [37] is used as a loss function. RMSprop is a gradient-based optimization technique used to change weights and the learning rate of the neural network to reduce the losses. We repeat the same process for MLP, GRU, and other traditional machine learning models. For each model, we note down each model's accuracy and F1-Score. Figure 3 shows the proposed architecture of the stylometric features for the traditional and deep learning models.

Models with TF-IDF Features. We implement traditional machine learning and deep learning models to identify gender from the text using Term Frequency Inverse Document Frequency (TF-IDF) features. We initialize the TF-IDF vectorizer with n-gram (n = 2) and max_feature set to 1000. Then we transform our data using the TF-IDF vectorizer. Figure 4 shows the architecture of traditional machine learning and deep learning models with TF-IDF features.

Models with Word Embedding Features. For any sentence S with classification C, we have done the necessary pre-processing as discussed in Section IV-A3. Then these sentences are passed through a tokenizer which can produce a one-hot encoding vector of length 100. Only the top 1000 most frequent words are taken as vocabulary. The first 100 words are taken for sentences having more than 100 words. Shorter text is padded with zeros. After that, these vectors are fed into an embedding layer. The weights of the embedding layer are initialized with word2vec embedding weights. We initialize the embedding layer using trainable and non-trainable properties. Non-trainable property freezes the Embedding layer so that the pre-trained weights are not updated during the training. The output dimension of the embedding layer is 300 as it is the vector length of each word in the word2vec model. The sequence of 100 words is then passed through an LSTM layer. The output of the LSTM layer is passed to a dense layer which is used to detect gender. Sigmoid [36] is used in the dense layer as an activation function. The optimizer is RMSprop and binary cross entropy [37] is used as a loss function. The same process is repeated for all the remaining deep-learning models. Figure 5 shows the architecture of our models with the word embedding features.

C. Experimental Setup

We use the Python Keras framework with Tensorflow to implement deep learning models for training, tuning, and testing. For the text classification using the relations between words in our dataset, we use the Gensim word2vec model. We use the scikit-learn library to implement traditional machine learning algorithms. Experimental evaluation was conducted on a machine with an Intel Core i7 processor with 1.8GHz clock speed and 8GB RAM. The machine has also an NVIDIA GeForce MX150 with 2GB memory and therefore Tensorflow-based experiments have fully utilized GPU instructions. Considerable speed can be achieved in Tensorflow-based experiments by adding a GPU as shown in [38].

D. Performance Evaluation and Parameter Tuning

We assess the efficiency and scalability of our methods by varying model architectures and feature sets. Performance is measured using word vectors generated from stylometric features, TF-IDF vectorizers, and word embeddings. We evaluate different word2vec models with both trainable and non-trainable embedding weights, utilizing the RMSprop optimizer and binary cross entropy loss function. For LSTM and GRU algorithms, we set epochs and batch size as key parameters. Training accuracy plateaued after a certain epoch, and validation loss increased with more epochs, indicating overfitting. To address this, we limited epochs to 5 and used a batch size of 32 in all experiments, employing 10-fold cross-validation.

E. Results

In this section, we evaluate the performance of our proposed methods for gender identification on our dataset of 15,000 samples. We compare the performance of deep learning algorithms with traditional machine learning algorithms which answers the RQ4.

In addition to that, we answer RQ5 by delving into the analysis of the results, beginning with an examination of the performance of word2vec models, followed by a comparison of different machine learning models, and finally, an exploration of why TF-IDF excels in gender identification from Bangla social media texts.

Word2vec Performance. We employed both trainable and non-trainable weights for the Continuous Bag of Words (CBOW) and Skip-Gram (SG) architectures of the word2vec model. Table II reveals that the CBOW model consistently outperforms SG, exhibiting superior accuracy and F1-Score. Notably, initializing the embedding layer's weights as trainable in the CBOW architecture proves more effective than using static weight initialization. This preference for CBOW can be attributed to the nature

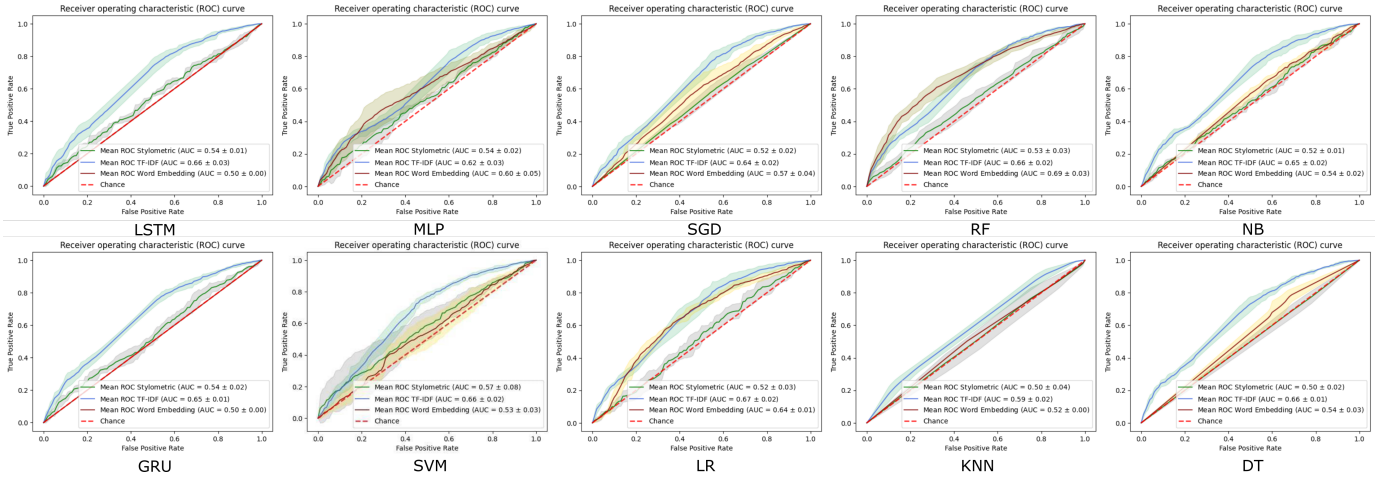

Fig. 6: Comparison of ROC curves among different machine learning models

	Accuracy(%)			F1-Score(%)		
Model	**Stylometric**	**TF-IDF**	**Word Embedding**	**Stylometric**	**TF-IDF**	**Word Embedding**
LST	74.91	73.46	75.62	80.23	84.09	81.21
GRU	74.91	73.71	73.98	80.24	84.02	79.92
MLI	70.82	76.67	72.94	77.28	76.98	84.09
SGD	75.64	**78.33**	76.02	75.00	**87.67**	86.38
RF	73.37	77.43	73.84	79.25	87.28	84.84
NB	63.02	77.82	74.23	68.34	87.18	85.18
SVM	72.68	77.46	74.41	78.85	86.93	84.84
LR	73.42	76.92	72.94	78.62	86.60	83.78
KNN	67.07	75.64	67.56	74.41	85.94	79.41
DT	63.24	73.07	70.38	69.17	82.95	80.57

TABLE I: Performance measure for deep learning vs traditional models

Vectorization	**Accuracy(%)**	**F1-Score(%)**
CBOW (T)	75.13	79.92
CBOW (NT)	75.62	81.21
Skip-Gram (T)	74.74	79.8
Skip-Gram (NT)	75.62	81.2

TABLE II: Performance measure for Word Vector Model. Here, T stands for trainable and NT for non-trainable.

of the dataset, which contains predominantly frequently occurring words rather than sparse or infrequent ones. CBOW predicts the most probable word (i.e., the most frequently occurring word) given a context, which aligns well with the dataset characteristics. Additionally, allowing the update of word embeddings in the embedding layer for SG can disrupt contextual information, favoring the non-trainable CBOW model.

Performance of Different Models. The performance of various machine learning models associated with different feature vectors is presented in Figure 6. Table I highlights the F1-Scores of traditional models (SGD, RF, and NB) surpassing those of deep learning models (LSTM and GRU) when using TF-IDF and Word Embedding features. This observation is expected as TF-IDF and Word Embeddings represent words numerically, with the former focusing on statistical word importance and the latter capturing semantic relationships. In contrast, deep learning models exhibit better performance with stylometric features, where GRU slightly outperforms LSTM. This aligns with the inherent capabilities of these features, where Word Embeddings capture context, and stylometric features capture authorial writing style. Additionally, SVM outperforms NB in terms of accuracy, as NB relies on a simpler text frequency computation to determine class probabilities. Notably, SGD stands out with the highest accuracy (77.43%) and F1-Score (87.67%) when utilizing TF-IDF features.

Deep Learning vs. Traditional Machine Learning Performance. The comparative analysis between traditional machine learning models and deep learning models, considering both traditional representations like TF-IDF and embedding representations like word2vec, is presented in Table I. Despite the curse of dimensionality associated with the Bag of Words (BOW) approach, TF-IDF proves effective in solving the gender classification problem. TF-IDF outperforms the context-based word2vec vectorizer approach, highlighting the power of traditional ML models over neural network models in this context. The widely adopted SGD algorithm demonstrates efficiency and success in author gender identification by creating linear decision boundaries. Conversely, neural network models like LSTM and GRU, which rely on weight-based categorization, exhibit underperformance in our study, consistent with previous research findings [14]. This can be attributed to the deep learning models' need for extensive datasets and their less effective capture of subtle stylistic variations critical for gender classification, a strength of the TF-IDF approach.

Understanding TF-IDF's Superiority.

In this section, we offer a thorough analysis to elucidate why TF-IDF surpasses word embeddings in gender identification from Bangla social media texts.

Discriminative Term Identification: TF-IDF excels in identifying discriminative terms within a corpus by

emphasizing terms that frequently occur in specific documents and are rare across the entire corpus. This property is valuable in gender identification, as it helps pinpoint terms characteristic of certain genders while differentiating them from common, non-discriminatory terms.

Stylistic Signifiers: Gender-specific writing styles often hinge on subtle lexical and stylistic choices. TF-IDF excels in isolating these stylistic signifiers by emphasizing terms carrying unique gender-related signals. This enables it to capture nuances of language use distinctive to each gender, such as preferred vocabulary, punctuation, and sentence structure.

Mitigating Common Term Overrepresentation: Social media texts often contain common terms and stopwords that do not significantly contribute to gender identification. TF-IDF mitigates their impact by assigning lower weights to them, ensuring a focus on salient linguistic features indicative of gender.

*1) **TF-IDF vs. Word Embeddings**:* In our study, TF-IDF outperformed word embedding techniques, such as CBOW and Skip-Gram models, due to several key factors. Word embeddings excel at capturing semantic and contextual word relationships, suitable for tasks like sentiment analysis or topic modeling. In contrast, gender identification relies on subtle stylistic variations in language use, where TF-IDF's emphasis on term frequency and distinctiveness aligns better. Moreover, word embeddings demand substantial training data, posing challenges in languages with limited data availability like Bangla, while TF-IDF relies on straightforward statistical measures, remaining effective with smaller datasets.

VI. Model Robustness

A. Testing in the Wild

To evaluate the real-world applicability of our classifier, we compiled a dataset consisting of live posts from well-known male and female profiles. This dataset included a total of 2700 posts, with 1270 sourced from male profiles and 1430 from female profiles. To assess the performance of our classifier, we varied the word counts in the posts, as detailed in Figure 7.

Our analysis revealed that the classifier achieved its highest accuracy of 63.21% at a word count of 70, and the best F1-score was 76.80% for posts with at least 100 words. On average, the classifier attained an accuracy of approximately 60.96% and an F1-score of 75.35%, indicating a relatively robust performance across different word count thresholds.

It is understandable that our prediction accuracy experienced a drop when tested with real-world posts. Given that we gathered a constrained dataset for our training, it's plausible that posts from the wider online community might contain numerous out-of-vocabulary words, affecting our model's performance.

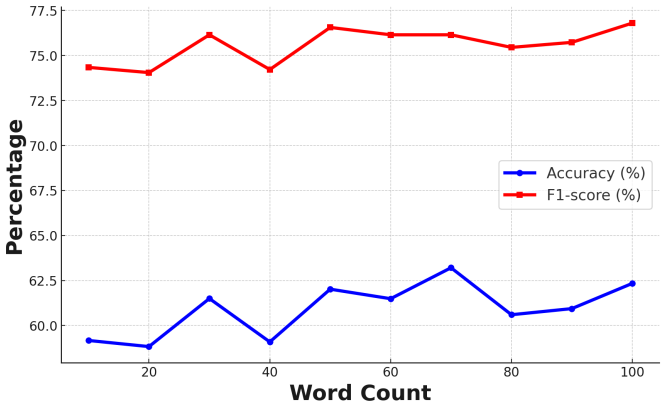

Fig. 7: Classifier's efficacy in real-world scenarios with varying word count.

B. Sensitivity Analysis

The robustness of our classification model to gender-specific "noise" was evaluated through a sensitivity analysis. We define the "noise" as labeling actual Male samples as Female and vice versa. This process involved incrementally mixing text data from the opposite gender into male and female datasets and observing the effects on model performance.

- **Mixing Proportions:** Starting with gender-specific datasets, we introduced text data from the opposite gender in 5% increments, varying from 5% to 50% of the dataset, to emulate the mixture of text data encountered in real-world scenarios.
- **Performance Metrics:** We calculated the model's accuracy, precision, recall, and F1 score at each increment to measure the impact of the gender-based "noise."

The sensitivity of our model was quantified by the changes in performance metrics as we introduced "noise" into the datasets. The results, depicted in Figure 8 and Figure 9, indicated that model accuracy and F1 score varied with the level of noise. A significant observation was that model accuracy declined by 4.75% when the noise reached 35%, and the highest F1 score was achieved when datasets contained 100 words.

Interpretation. Our findings from the sensitivity analysis suggest several key points:

- The model displayed resilience up to a threshold level of noise but began to lose reliability beyond a 30% mixture of opposite gender data.
- This threshold indicates a critical point where the model's performance is notably compromised.
- Identifying the points of performance decline provides insight into potential areas for model improvement, suggesting that our approach could benefit from incorporating noise management techniques.

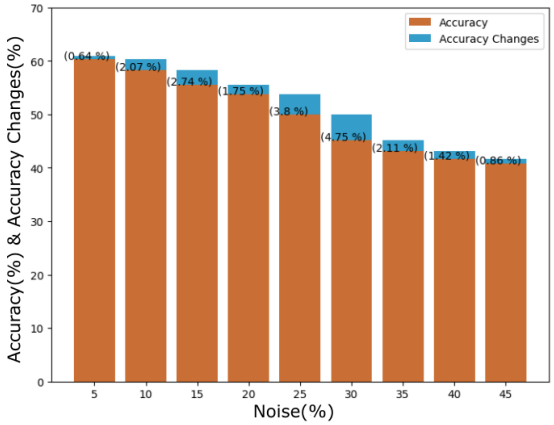

Fig. 8: Sensitivity of accuracy against the noise.

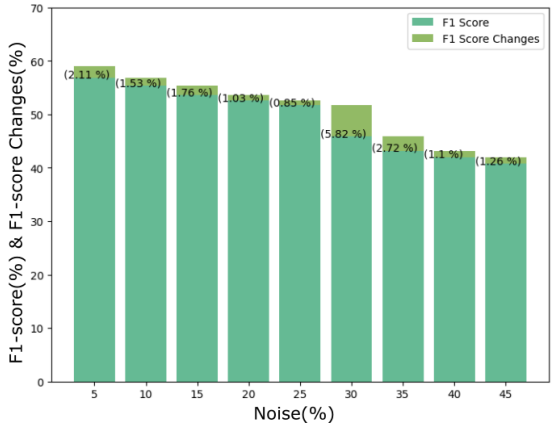

Fig. 9: Sensitivity of F1 score against the noise.

VII. Ethical Considerations

In our approach, which involved collecting data from public groups, profiles and pages, and other online sources, we strictly adhered to ethical guidelines and standards. Our protocols for data interaction and collection were thoroughly reviewed and approved by an Institutional Review Board (IRB). We ensured that only anonymized data was stored for analysis, aligning with the NIST SP 800-122 guidelines on Personally Identifiable Information (PII) [39]. This meant that any identifying details were carefully omitted from our dataset. Furthermore, in compliance with GDPR [40], our use of data, stripped of personal context like names or identification numbers, does not qualify as "personal information". This rigorous approach to data anonymization demonstrates our commitment to maintaining user privacy and ethical research practices.

VIII. Discussion and Limitations

Our study acknowledges several limitations and ethical considerations. With a dataset of only 15,000, it may not represent all Bangla-speaking social media users. The methodologies developed are context-specific and may not be transferable to other languages or scenarios. A key concern is the potential misuse of our gender prediction classifier, which could reinforce gender stereotypes or be used unethically. To mitigate these risks, we have implemented rigorous data anonymization, responsible data storage practices, and stringent access controls. Transparency is maintained through detailed method documentation, ensuring reproducibility. We strictly adhered to ethical research guidelines, underscoring the integrity of our work.

Future research in Bangla text-based gender identification is promising. Potential directions include expanding the dataset, testing TF-IDF in other regional languages, exploring hybrid models combining TF-IDF with deep learning, and applying TF-IDF to diverse text classification tasks. Additionally, leveraging the latest Large Language Models could further enhance gender identification capabilities.

IX. Conclusion

This research marks a significant advancement in NLP, specifically in gender identification from Bangla text, a previously uncharted area. Our approach combined classical and deep learning techniques to effectively determine gender from Bangla texts, using a diverse collection of online texts for model training and evaluation. Comparative assessments with similar studies in Arabic, Russian, and English revealed that classical machine learning models surpass deep learning models in accuracy and F1-Score for Bangla texts, highlighting the relevance of traditional methods in language-specific NLP tasks.

X. Acknowledgment

This research was supported by NSF grant CNS-2153482.

References

[1] Y. Joo, I. Hwang, L. Cappellato, N. Ferro, D. Losada, and H. Müller, "Author profiling on social media: An ensemble learning model using various features," *Notebook for PAN at CLEF*, vol. 2380, 2019.

[2] E. Min, Y. Rong, Y. Bian, T. Xu, P. Zhao, J. Huang, and S. Ananiadou, "Divide-and-conquer: Post-user interaction network for fake news detection on social media," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 1148–1158.

[3] C. Deutsch and I. Paraboni, "Authorship attribution using author profiling classifiers," *Natural Language Engineering*, vol. 29, no. 1, pp. 110–137, 2023.

[4] S. Talukder and B. Carbunar, "A study of friend abuse perception in facebook," *ACM transactions on social computing*, vol. 3, no. 4, pp. 1–34, 2020.

[5] G. J. Park, D. Yaden, H. A. Schwartz, M. L. Kern, J. C. Eichstaedt, M. Kosinski, D. Stillwell, L. H. Ungar, and M. E. Seligman, "Women are warmer but no less assertive than men: Gender and language on facebook," *PLoS ONE*, vol. 11, 2016. [Online]. Available: https://consensus.app/papers/women-warmer-less-assertive-gender-language-facebook-park/36174b02201d55099409cd54a89721d1/?utm_source=chatgpt

[6] M. H. Klaiman and A. Lahiri, "Bengali," in *The world's major languages*. Routledge, 2018, pp. 427–446.

[7] Chung Hwan Kwak, "New world encyclopedia," https://www.newworldencyclopedia.org/entry/Bengali_language, 2020, [Online: accessed 09-April-2023].

[8] Simon Kemp, "Digital 2021: Bangladesh," https://datareportal.com/reports/digital-2021-bangladesh, 2021, [Online: accessed 17-Jun-2021].

[9] StatCounter Global Stats, "Social media stats in bangladesh," https://gs.statcounter.com/social-media-stats/all/bangladesh, 2020, [Online: accessed 24-May-2021].

[10] S. Ahmed, M. J. Alam, S. Talukder, and I. Hossain, "Towards addressing identity deception in social media using bangla text-based gender identification," in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 2023, pp. 72–76.

[11] M. Corney, O. d. Vel, A. Anderson, and G. Mohay, "Gender-preferential text mining of e-mail discourse," in *Computer Security Applications Conference(CSAC), Las Vegas, USA, Dec 9-13*, 2002.

[12] C. R. Fink, D. S. Chou, J. J. Kopecky, and A. J. Llorens, "Coarse- and fine-grained sentiment analysis of social media text," *Johns Hopkins APL Technical Digest*, vol. 30, no. 1, pp. 22–30, 2011.

[13] M. B. o. Vollenbroek, T. Carlotto, T. Kreutz, M. Medvedeva, C. Pool, J. Bjerva, H. Haagsma, and M. Nissim, "Content-centric age and gender profiling," in *Notebook for PAN at CLEF, Évora, Portugal, Sept 05-08*, 2016.

[14] A. Sboev, T. Litvinova, I. Voronina, D. Gudovskikh, and R. Rybka, "Deep learning network models to categorize texts according to author's gender and to identify text sentiment," in *Computational Science and Computational Intelligence (CSCI), Las Vegas, USA, Dec 15-17*, 2016.

[15] H. Mubarak, S. A. Chowdhury, and F. Alam, "Arabgend: Gender analysis and inference on arabic twitter," *arXiv preprint arXiv:2203.00271*, 2022.

[16] K. Alsmearat, M. Al-Ayyouba, R. Al-Shalabi, and G. Kanaanbt, "Author gender identification from arabic text," *Journal of Information Security and Applications*, vol. 35, no. 8, pp. 85–95, 2017.

[17] B. Bsir and M. Zrigui, "Enhancing deep learning gender identification with gated recurrent units architecture in social text," *Computación y Sistemas*, vol. 22, no. 3, pp. 757–766, 2018.

[18] N. Cheng, R. Chandramouli, and K. Subbalakshmi, "Author gender identification from text," *Digital investigation*, vol. 8, no. 1, pp. 78–88, 2011.

[19] S. Rahman, F. Kabir, and M. N. Huda, "Automatic gender identification system for bengali speech," in *Electrical Information and Communication Technologies (EICT), Khulna, Bangladesh, Dec 10-12*, 2015.

[20] N. I. Tripto and M. E. Ali, "Detecting multilabel sentiment and emotions from bangla youtube comments, sylhet, bangladesh," in *International Conference on Bangla Speech and Language Processing (ICBSLP), Sept 21-22*, 2018.

[21] Z. M. Nia, A. Ahmadi, B. Mellado, J. Wu, J. Orbinski, A. Agary, and J. D. Kong, "Twitter-based gender recognition using transformers," *arXiv preprint arXiv:2205.06801*, 2022.

[22] M. Simmons and J. S. Lee, "Catfishing: A look into online dating and impersonation," in *Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis: 12th International Conference, SCSM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22*. Springer, 2020, pp. 349–358.

[23] D. Lee, "Facebook, Twitter and Google berated by senators on Russia," [BBC Technology] tinyurl.com/ybmd55js, 2017.

[24] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. Kelley, D. Kumar, D. McCoy, S. Meiklejohn, T. Ristenpart, and G. Stringhini, "Sok: Hate, harassment, and the changing landscape of online abuse," in *IEEE Symposium on Security and Privacy (SP)*, 2021, pp. 473–493.

[25] E. M. Redmiles, N. Chachra, and B. Waismeyer, "Examining the demand for spam: Who clicks?" in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 212:1–212:10.

[26] ——, "Examining the demand for spam: Who clicks?" in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–10.

[27] Mulac, A., "The gender-linked language effect: do language differences really make a difference?" https://psycnet.apa.org/record/2006-03342-012, 2021, [September 09, 2021].

[28] Luthfi Ramadhan, "Tf-idf simplified," https://towardsdatascience.com/tf-idf-simplified-aba19d5f5530, 2020, [Online: accessed 30-October-2022].

[29] S. Alam, M. A. U. Haque, and A. Rahman, "Bengali text categorization based on deep hybrid cnn–lstm network with word embedding," in *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*. IEEE, 2022, pp. 577–582.

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *arXiv preprint arXiv:1301.3781*, 2013.

[31] A. Ahmad and M. R. Amin, "Bengali word embeddings and it's application in solving document classification problem," in *2016 19th international conference on computer and information technology (ICCIT)*. IEEE, 2016, pp. 425–430.

[32] I. Hossain, S. Puppala, M. J. Alam, and S. Talukder, "Monitoring dynamics of emotional sentiment in social network commentaries," in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 2023, pp. 51–55.

[33] ——, "A visual approach to tracking emotional sentiment dynamics in social network commentaries," 2024.

[34] stopwords-iso, "Stopwords bengali," https://github.com/stopwords-iso/stopwords-bn, 2021, [Online: accessed 31-May-2021].

[35] S. Brody and N. Diakopoulos, "Cooooooooooooooollllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs," in *Proceedings of the 2011 conference on empirical methods in natural language processing*, 2011, pp. 562–570.

[36] O. Sharma, "A new activation function for deep neural network, faridabad, india," in *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMIT-Con), Feb 14-16*, 2019.

[37] Shipra Saxena, "Binary cross entropy/log loss for binary classification," https://www.analyticsvidhya.com/blog/2021/03/binary-cross-entropy-log-loss-for-binary-classification/, 2021, [MARCH 3, 2021].

[38] M. et al., "Gpflow: A gaussian process library using tensorflow," *Journal of Machine Learning Research*, vol. 18, no. 40, pp. 1–6, 2017. [Online]. Available: http://jmlr.org/papers/v18/16-537.html

[39] "Guide to protecting the confidentiality of personally identifiable information (pii)," https://tinyurl.com/ylyjst5y, 2021, accessed: 2023-02-12.

[40] "General data protection regulation (gdpr)," https://gdpr-info.eu/, 2021, accessed: 2023-02-12.