Towards Addressing Identity Deception in Social Media using Bangla Text-Based Gender Identification

Sultan Ahmed*, Md Jahangir Alam†, Sajedul Talukder‡ and Ismail Hossain§
*University of Maryland, Baltimore County, USA
Southern Illinois University, Carbondale, USA
Email: *sultan.ahmed@umbc.edu, †mdjahangir.alam@siu.edu, ‡sajedul.talukder@siu.edu, §ismail.hossain@siu.edu

***Abstract*—Gender identification from social media content can play a crucial role in detecting and mitigating the risks posed by counterfeit accounts. Authentic gender representation can foster a safer and more diverse online environment. While research has been conducted on gender identification in languages such as English, Russian, and Arabic, there remains a gap in studies targeting Bangla and its related languages. This paper introduces a stylometric feature approach to discern the gender of authors from Bangla texts. Utilizing a dataset of 5,000 posts sourced from various Facebook groups, we trained seven traditional machine learning models. Among these, the Random Forest (RF) model notably excelled, achieving an accuracy of 73.37% and an F1-Score of 79.25%, thus setting a promising benchmark in gender identification from Bangla texts.**

***Index Terms*—social network friend spam, gender identification, Bangla language, stylometric features, word vector**

I. Introduction

The absence of authenticated user data on online platforms, notably social networks, paves the way for the surge of counterfeit and anonymous profiles. Such issues could be mitigated with Author Profiling (AP) [10]. Bangla, ranking as the fifth predominant native Indo-European language and seventh globally, encompasses approximately 300 million native and an additional 37 million secondary speakers [11], [4].

Serving as the primary language for 98% of the Bangladeshi populace, Bangla is formally recognized as the national language of Bangladesh [6]. Its influence is not restricted to national boundaries, with significant Bangla diaspora in regions such as the Middle East, Europe, and the USA [4]. Amidst the emergence of a Digital Bangladesh initiative [19], Bangla's digital imprint on platforms like Facebook, Linkedin, and Twitter has grown. Facebook, emerging as Bangladesh's dominant social platform [20], claims 33.71 million active Bengali users [13]. Engaging in commerce, communication, and more via Facebook groups and Messenger, the urgency to address the proliferation of inauthentic Facebook profiles, especially concerning gender and age, is evident. This manuscript delves into gender identification from Bangla text on social media. With exhaustive research in languages like English, Russian, and Arabic, the dearth of work in Bangla underscores the significance of our endeavor.

Facebook's textual content, inherently conversational, often conveys pivotal insights and viewpoints. Gender identification automation from such content necessitates grasping both writing style and underlying context, going beyond rudimentary rule-based approaches. An author's writing style mirrors their psychological, sociological, and even physiological states. Stylometric features (SF) delineate unique aspects of this style, encompassing elements such as word frequency, sentence length, or special character utilization. Consider the sentence, "সে এভাবেই লিখে যেতে থাকে! (He continues to write in this manner)": stylometric attributes include 22 characters, 5 words, 4 spaces, and 1 special character. The core hypothesis of this paper asserts that every author, or author groups with shared attributes like gender or age, exhibit distinctive, relatively static writing styles. Leveraging this consistency allows for precise author identification, enabling insights into aspects like age, gender, or psychological states using stylometric features.

In this manuscript, we introduce machine learning techniques tailored for gender identification derived from

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11…$15.00
https://doi.org/10.1145/3625007.3627732

Language	Feature	Model
Arabic [1]	Character based, word based, syntactic, semantic functional word	BN, NB, Logis, ANN, SGD, SLogis, SVM, KNN, DT, JRip, OneR, PART, J48, RF (Best accuracy: 80.4%)
Arabic [2]	Lexical features, n-grams, syntax features, character frequencies, Bag of smileys, Arabic stop words	Bi-Gated Recurrent Unit Layer (Best accuracy: 79%)
English [3]	Character-based, Word-based, Syntactic, and Structural features, Function words	SVM, Adaboost Decision Tree, Logistic Regression (Best accuracy: 83%)
Russian [18]	Morphological features, Syntactical parameters, Derivative coefficients, No. of words pertaining emotion	Gradient Boosting, Adaptive Boosting, ExtraTrees, Random Forest,PNN, Support Vector Machine, CNN, LSTM (Best accuracy: 77%)
Bangla (this work)	Character based, word based, syntactic, semantic, functional word	RF, NB, SVM, LR, KNN DT (Best accuracy: 73.37%)

TABLE I: Comparative analysis of methods done in various languages with their performances.

textual data. Leveraging a dataset collated from various Facebook group posts, our primary objective is to evaluate the efficacy of our methods specifically within the context of the Bangla language. While numerous traditional machine learning algorithms (that eschew deep learning paradigms) have been tested for performance comparison, it's noteworthy to mention that, to our comprehension, no extant research has ventured into gender identification using Bangla datasets. The salient contributions of this paper encapsulate:

- The formulation of a bespoke dataset comprised of 5000 samples, amalgamating user-generated text with corresponding gender data in Bangla.
- The introduction of an array of stylometric features, envisaged as potent gender determinants for the Bangla lexicon.
- The conceptualization and execution of a comprehensive experimental framework, aiming to deduce gender from succinct textual samples.

II. Related Works

Initial investigation in identifying the gender of the author was done from an email corpus with different features like structural and gender preferential features [5] using Support Vector Machine (SVM) with a maximum average accuracy of 71.2%. Experiments to automatically detect users' gender, age, mental, social state, etc. are conducted widely on Facebook, Twitter [23], Blog [17] and telephone conversation [16] data. Fink et al. [8] found sentiment through analysis of Facebook text data using Naive Bayes and Support Vector Machine algorithms. Stylistic features like capitalized letters, tokens, punctuation, function words, parts of speech, etc. have been employed in English and Spanish [22] language for age and gender profiling with an average accuracy of 52.58% for English. Content-based features were observed dominant than other features in both English and Spanish datasets.

Another line of work has focused on inferring gender by utilizing usernames, profile pictures, tweets, and networks of friends from Arabic Twitter-sphere [14], identifying genders from the Arabic data based on articles from online newspaper websites [1], classifying gender the gender of Arabic, German, Iranian and Japanese first names [12], and differentiating male and female authors of tweets in Egyptian Arabic dialect [7] and identifying gender in Egyptian tweets [9]. Other studies have explored the use of NLP techniques and tweets in a gender classification system [21]. Cheng et al. [3] detected genders from English datasets (institutional emails and online newspapers) using stylometric features with traditional machine learning algorithms, including SVM, AdaBoost Decision Tree, and logistic regression with the average accuracy of 83%, 75%, and 73% respectively. Cheng et al. [3] identified genders from the English dataset including emails of institutions and the online newspaper. They applied stylometric features in traditional machine learning algorithms such as SVM, AdaBoost Decision Tree, and logistic regression. Table I shows a comprehensive comparison of performance of gender identification in different languages and Bangla language.

III. Dataset

A. Data Collection

To create our dataset, we extract Facebook posts with author names from three different public Facebook groups related to the Bangladesh University of Engineering and Technology. To ensure our data consists of posts from authenticated members, some exclusion criteria were applied during the search phase. First, we screened the Facebook user's name to ensure it sounds legit. If eligibility could not be determined based on the name, we retrieved the full text to confirm inclusion or note reasons for exclusion. Steps were then taken to de-identify users' personal or written (i.e., photo) information attached to the

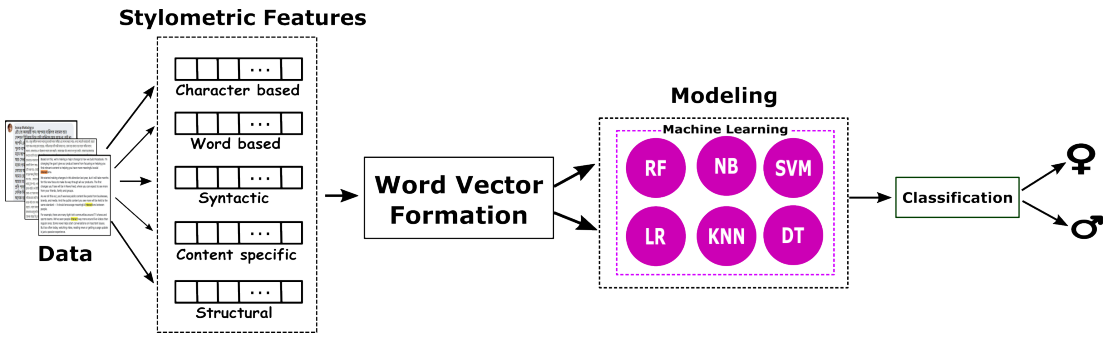

Fig. 1: Model Architecture for Stylometric Feature

Group Name	Group Link (Currently Private)
1. বুয়েটে আড়িপেতে শোনা	OverheardAtBUET
2. BCS Preparation for BUET Students	BCS.Aspirants.BUETian
3. এই পাপী চোখে বুয়েটে যা দেখেছি	895957014215386 (Archived)

TABLE II: Facebook Groups for Data Collection

user. We manually label the members as Male and Female. Then we assign the author's gender by cross-multiplying the Facebook post data with the author's name and labeled members' list. Thus a dataset containing authenticated user posts with gender labeling is created. We applied a cryptographic hash function to the Facebook account names and used the anonymized names to store our data. The group names are listed in Table II. We collect 5000 posts in total out of which 3035 are from male authors and 1965 are from female authors. The gender split of the Facebook posts with author names was in line with that of the student body at the Bangladesh University of Engineering and Technology, where there are fewer female students (20%-30%) than male students. To reduce the data imbalance, we manually collect some posts from the Facebook profile of the female members.

B. Pre-processing

The language used in member posts on any Facebook group is often informal. Here, the text contains URLs, images, tags, links, etc. So, before training, we use a variety of data pre-processing techniques. All characters other than Bangla alphanumeric letters, punctuation, and URLs are removed during the pre-processing step. We also delete the images, links, hashtags, and user tags. Stop words frequently include extraneous details and repetition. As a result, we tokenize our texts and eliminate stop words.

C. Ethical Considerations

We have developed our protocols to interact with Facebook and collect data in an ethical, IRB-approved manner. For analysis purposes, we only stored anonymized data that we extracted from Facebook users and groups.

IV. Methodology

In the following, we discuss the detailed methodology of our research. We first create a dataset consisting of Facebook posts collected from different Facebook groups and annotate the data before performing the feature extraction. Then, to reduce noise from the data, we use a variety of pre-processing algorithms on the text. To convert each text to a particular input format, we implement the stylometric feature approach for feature extraction. Finally, we present the architecture of our model.

A. Word Vector Formation

We use deep learning models in our proposed solution. To use the deep learning models, we convert our text to a word vector using stylometric feature approach.

*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 [5]. These features are categorized into four types: lexical features, structural features, syntactic features, and content-specific features.

Lexical features 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 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 man's and woman's different habits of using punctuation. For example, women tend to use more question marks than men [15].

Structure-based features 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

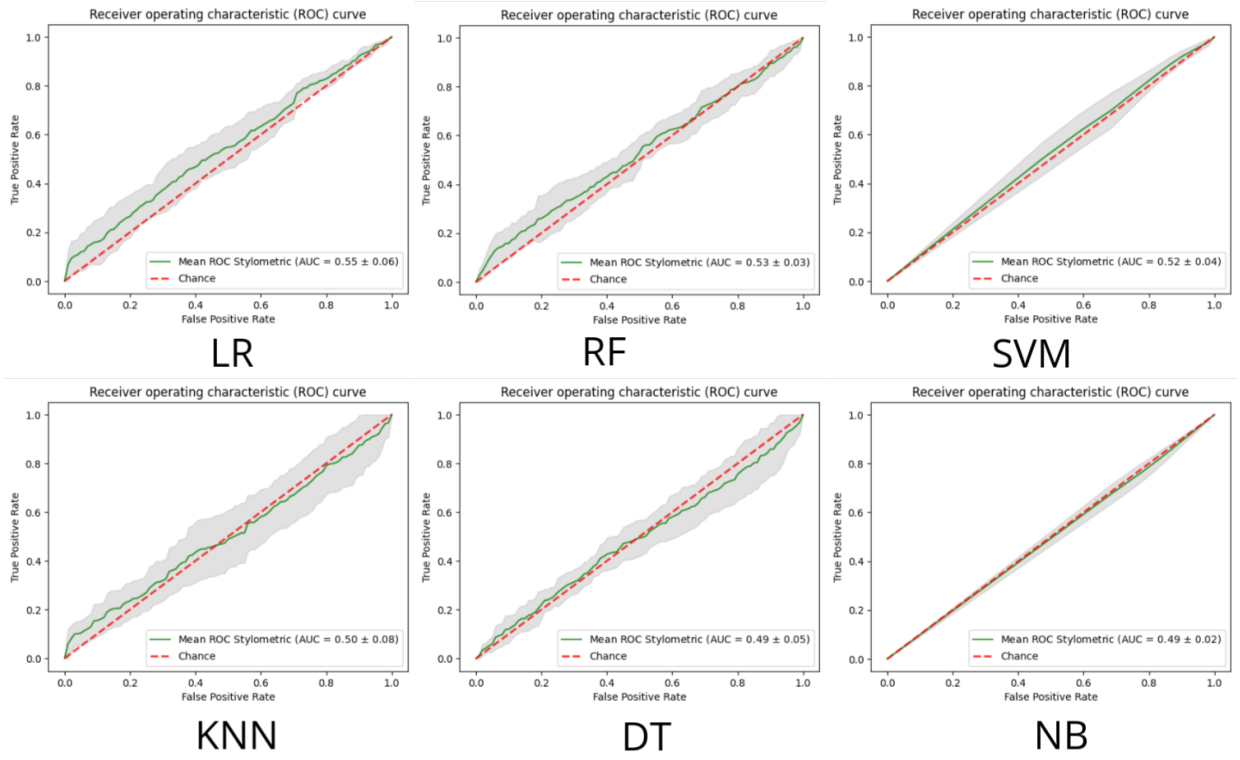

Fig. 2: Comparison of ROC curves among different traditional machine learning models.

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 represent domain-specific terms. For these features, we first collect the feature words as suggested for the Arabic language in [1]. 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:

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

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.

V. Model Architecture

We implement traditional machine learning model architecture for gender identification. In this architecture, we implement Support Vector Machine, Naive Bayes, K-Nearest Neighbors, Random Forest, Decision Tree and Logistic Regression models. Figure 1 shows the architecture of our proposed model.

A. Models with Stylometric Features

We prepare word vectors using the stylometric feature approach discussed in Section IV-A1. Then the word vector is fed into traditional machine learning models. We train models with default parameters of each algorithm. For each model, we note down model accuracy and F1-Score which are shown in Table III. Figure 1 shows the proposed architecture of gender identification for stylometric features using the traditional models.

VI. Results

In this section, we evaluate the performance of our proposed methods for gender identification on the Facebook dataset. We compare the performance of different traditional machine learning algorithms.

Experimental Setup We use the Python Scikit-learn library to implement traditional machine learning models for training, tuning, and testing. Experimental evaluation was

conducted on a machine with an Intel Core i7 processor with 1.8GHz clock speed and 8GB RAM.

Performance Evaluation We study the efficiency and scalability of our proposed methods by varying model architectures and feature sets. We measure the performance of our models by generating word vectors using the stylometric feature.

A. Result Analysis

*1) **Performance of different models.**:* We present the performance of different machine learning models associated with the feature vector in Figure 2. We measure the model performance by observing the F1-score and accuracy. We list the F1-score and accuracy of each model in Table III. From the table we can see that the F1-Score of 79.25% of RF outperforms the other models.

Model	Accuracy	F1-Score
RF	73.37%	79.25%
NB	63.02%	68.34%
SVM	72.68%	78.85%
LR	73.42%	78.62%
KNN	67.07%	74.41%
DT	63.24%	69.17%

TABLE III: Performance measure for machine learning models

VII. Conclusion & Future Work

Since no other research has been done to determine gender from Bangla text, gender identification from Bangla text represents the state of the art in NLP. In this study, we have constructed a model to extract gender information from Bangla text using traditional machine learning techniques. We have gathered a text dataset from different Facebook groups. Then, our model is assessed using this data set. Additionally, we evaluated the efficacy of our model against studies on the Arabic, Russian, and English languages. In terms of accuracy and F1-Score for the Bangla language, the classical model surpasses the deep learning model. In subsequent research, we will expand our decision engine to recognize additional author characteristics including age, background, native tongue, political views, etc. Based on the writing style, authorship will be assigned using the current decision engine. To improve the deep learning model's performance, the data set can be improved. Other languages can use the strategies we discussed in this paper.

VIII. Acknowledgment

This research was supported by NSF grant CNS-2153482.

References

[1] K. Alsmearat, M. Al-Ayyouba, R. Al-Shalabi, and G. Kanaanbt. Author gender identification from arabic text. Journal of Information Security and Applications, 35(8):85--95, 2017.

[2] B. Bsir and M. Zrigui. Enhancing deep learning gender identification with gated recurrent units architecture in social text. Computaci□n y Sistemas, 22(3):757--766, 2018.

[3] Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. Author gender identification from text. Digital investigation, 8(1):78--88, 2011.

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

[5] M. Corney, O. de 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.

[6] Eglitis-media. worlddata.info. https://www.worlddata.info/languages/bengali.php, 2020. [Online: accessed 09-April-2023].

[7] Shereen ElSayed and Mona Farouk. Gender identification for egyptian arabic dialect in twitter using deep learning models. Egyptian Informatics Journal, 21(3):159--167, 2020.

[8] 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, 30(1):22--30, 2011.

[9] Shereen Hussein, Mona Farouk, and ElSayed Hemayed. Gender identification of egyptian dialect in twitter. Egyptian Informatics Journal, 20(2):109--116, 2019.

[10] Youngjun Joo, Inchon Hwang, L Cappellato, N Ferro, D Losada, and H Mu□ller. Author profiling on social media: An ensemble learning model using various features. Notebook for PAN at CLEF, 2380, 2019.

[11] Miriam Holly Klaiman and Aditi Lahiri. Bengali. In The world's major languages, pages 427--446. Routledge, 2018.

[12] Shervin Malmasi. A data-driven approach to studying given names and their gender and ethnicity associations. In Proceedings of the Australasian Language Technology Association Workshop 2014, pages 145--149, 2014.

[13] Miniwatts Marketing Group. Internet 2021 usage in asia. https://www.internetworldstats.com/stats3.htm, 2021. [Online: accessed 24-May-2021].

[14] Hamdy Mubarak, Shammur Absar Chowdhury, and Firoj Alam. Arabgend: Gender analysis and inference on arabic twitter. arXiv preprint arXiv:2203.00271, 2022.

[15] 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].

[16] F. Rangel, P. Rosso, M. Potthast, and B. Stein. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. In Notebook for PAN at CLEF, Dublin, Ireland, Sept 11-14, 2017.

[17] S. Rosenthal, A. Agarwal, and K. McKeown. Columbia nlp: Sentiment detection of sentences and subjective phrases in social media. In Semantic Evaluation (SemEval), Dublin, Ireland, Aug 23-24, 2014.

[18] 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.

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

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

[21] Pradeep Vashisth and Kevin Meehan. Gender classification using twitter text data. In 2020 31st Irish Signals and Systems Conference (ISSC), pages 1--6. IEEE, 2020.

[22] 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, □vora, Portugal, Sept 05-08, 2016.

[23] A. Zouaghi, L. Merhbene, and M. Zrigui. Combination of information retrieval methods with lesk algorithm for arabic word sense disambiguation. Artificial Intelligence Review, 38(4):257--269, 2012.