An Empiric Study on Bangla Sentiment Analysis Using Hybrid Feature Extraction Techniques

Md. Shymon Islam
Computer Science and Engineering Discipline
Khulna University
Khulna 9208, Bangladesh
Email: shymum1702@cseku.ac.bd

Kazi Masudul Alam
Computer Science and Engineering Discipline
Khulna University
Khulna 9208, Bangladesh
Email: kazi@cseku.ac.bd

***Abstract*—In this modern technologically advanced world, Sentiment Analysis (SA) is a very important topic in every language due to its various trendy applications. But SA in Bangla language is still in a dearth level. This work focuses on examining different hybrid feature extraction techniques on Bangla SA using a new comprehensive dataset of 203,493 comments collected from various microblogging sites. In this study, we have implemented 21 different hybrid feature extraction methods including Bag of Words (BOW), N-gram, TF-IDF, TF-IDF-ICF, Word2Vec, FastText, GloVe, Bangla-BERT etc with CBOW and Skipgram mechanisms. The proposed novel method (Bangla-BERT+Skipgram) outperforms all other feature extraction techniques in machine leaning (ML), ensemble learning (EL) and deep learning (DL) approaches. The (Bangla-BERT+Skipgram) method achieved highest 92.37%, 92.55% and 95.71% accuracy in ML, EL and DL algorithms.**

***Index Terms*—Sentiment Analysis, TF-IDF-ICF, FastText, Bangla-BERT, Skipgram**

I. Introduction

Sentiment analysis (SA) is a process of determining a person's views on a particular topic [1]. SA is the mining study of human opinion that analyzes people's opinions and feelings towards social entities such as services and products [2]. SA classifies the polarity of a document as whether the opinion being communicated through reviews and comment are positive, negative, or neutral [3-4]. The task of correctly identifying the polarity of a document to some predefined categories basically termed as sentiment analysis [5].

In this century, one cannot imagine a day without using the internet [6]. Millions of posts, blogs and comments are gathered on the internet everyday. So, microblogging sites become a great source of user generated Bangla reviews and comments [7]. Numerous research studies have been done on SA in English, Chinese, Hindi, Japanese, Arabic and Urdu languages [7-8] while sentiment analysis in Bangla language is still in a dearth level [9] though Bangla is the seventh most spoken language and is used daily by more than 250 million people in the world [10]. SA have many trendy applications such as book review analysis [1], online food delivery reviews analysis [3], conversation review analysis [5], product review analysis [8], social media monitoring [9], aspect based analysis [11], movie review analysis [12], market trend analysis [13] etc.

SA is an application of supervised learning algorithms, so feature extraction techniques are required. This work focuses on examining hybrid feature extraction techniques on Bangla SA using a new comprehensive dataset along with machine learning(ML), ensemble learning(EL), and deep learning(DL) algorithms. The followings are the contributions of the proposed work:

1) A newly created comprehensive Bangla sentiment corpus of 203,463 comments from 5 microblogging sites (Facebook, YouTube, Instagram, TikTok, Likee), manually tagging them into 15 categories containing 204,6150 tokens and 164,974 unique tokens.
2) Validate the dataset by 40 native Bangla speakers with a validation accuracy of 94.67%.
3) Examining various feature extraction techniques such as BOW, N-gram, TF-IDF, TF-IDF-ICF, Word2Vec, FastText, GloVe, Bangla-BERT etc and make a comparative study among them.
4) The proposed novel hybrid feature extraction (BERT+Skipgram) method outperforms all other techniques.
5) Applying ML, EL and DL algorithms that generates better performance in different metrics compared to state-of-the-art techniques.

The organization of this article: Section 2 describes the related works, Section 3 is for proposed methodology, Section 4 demonstrates the experimental results analysis and discussions and Section 5 concludes the works with some future remarks.

II. Related Works

Some of the recent studies of Bangla SA are discussed and summarized here.

In 2023, Kabir et al. [1] proposed a new large-scale Bangla dataset (BanglaBook) collected from book reviews that consists of 158,065 samples. They used BOW and N-grams as feature extraction methods, afterwards implemented ML and DL classifiers and obtained highest 0.9331 f1-score with BERT classifier. Another work from

Nafisa et al. [2] in the year of 2023, compiled a method for bipolar SA of online news comments, implemented six ML models along with BOW and TF-IDF transformers and a DL approach LSTM along with Word2Vec metric. They obtained highest 80% accuracy with RF and 83% with LSTM. In 2023, another new study was performed by Bitto et al. [3] for the user reviews collected from food delivery startups. They collected 1400 reviews from 4 food delivery Facebook pages and applied bipolar SA. Applying ML and DL algorithms, they obtained highest accuracy 89.64% using XGB and 91.07% from LSTM classifier. In 2022, Bhowmik et al. [4] proposed a method using an extended lexicon dictionary and DL algorithms deployed in two aspect based datasets collected from [11]. They obtained highest 84.18% accuracy using a hybrid model BERT-LSTM. At [5] the authors proposed a method for Bangla conversation reviews in 2022, they collected 1141 data from Bangla movies and short film scripts and implemented seven ML algorithms and recorded highest 85.59% accuracy with SVM. Another study carried by Prottasha et al. [6] in 2022 focused on transfer learning strategy of BERT based supervised fine tuning. They examined 6 different publicly available datasets and obtained highest results with the hybrid model CNN-BiLSTM. Another DL based study was done in [7] using LSTM, GRU and BLSTM classifiers along with 10-fold cross validation and achieved highest 78.41% accuracy score.

For product review mining, Shafin et al. [8] proposed a method in 2020 collecting 1020 reviews from Bangla e-commerce sites and examined five ML algorithms along with TF-IDF and recorded highest 88.91% accuracy with SVM. In [9], the authors made a dataset of only 7500 sentences and tested the trained models with only 100 sentences and achieved an accuracy above 90%. Two aspect based SA datasets were proposed in [11], Cricket (2900 samples) and Restaurant (2600 samples) datasets. Alam et al. [12] proposed a balanced dataset of 120000 documents and BRBT dataset (including Bangla and Romanized data) was proposed in [13] of 9337 samples. Compared to the literature being studied [1-13] for Bangla SA, it is proved that a rich dataset is always a crisis. To the best of our knowledge, we have created one of the largest document-level Bangla SA corpus of 203,463 comments from social media. The development of (Bangla-BERT+Skipgram) is the novelty of our proposed method.

III. Proposed Methodology

In this work, we have proposed a method for Bangla SA using a new comprehensive dataset and applying various hybrid feature extraction techniques. Figure 1 illustrates the workflow of the proposed method.

A. Data Collection

We have collected Bangla comments by using our self-developed crawlers from 5 micro-blogging sites, a total of 203,463 Bangla comments were collected and saved in an excel file. The overall summary of the domain based data collection is presented in Table I.

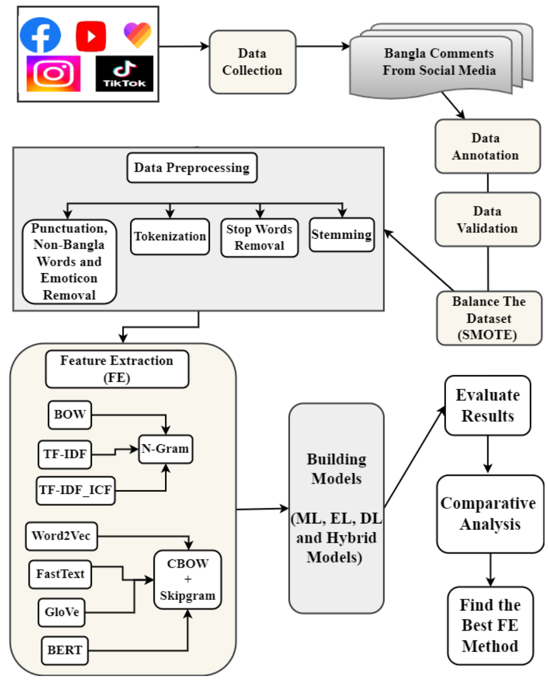

Figure 1. Workflow of the Proposed Bangla SA System

Table I
Summary of Domain Based Data Collection

Data Source	No. of Comments	Collection Period
Facebook	71,429	2022-2023
YouTube	42,884	2022-2023
Instagram	12,764	2022-2023
TikTok	52,367	2022-2023
Likee	24,019	2022-2023

B. Data Annotation and Validation

We have annotated the collected 203,463 Bangla comments manually by human experts (ourselves) into 15 basic sentiment categories which takes around six month. Among the 15 categories, the base emotion classes are subsequently mapped to 3 higher sentiment categories. To test the validity of the annotation process, we have conducted an analysis with 40 native Bangla speakers who are graduate students of North Western University, Bangladesh, each student was given 100 different comments and asked to annotate them into 15 predefined sentiment categories. So after running an audit on a total of 4000 comments (with 800 samples each from Facebook, YouTube, Instagram, TikTok and Likee) we found that it provides 94.67% accuracy. The confusion matrix of the annotation process is shown in Figure 2 and the complete description of the dataset is given in TABLE II.

Table II
OVERVIEW OF CATEGORY-WISE DATA COLLECTION

Higher Category	Basic Category	No. of Total Comments	No. of Total Tokens	No. of Unique Tokens	Topic Keywords (Top 2)	No. of Higher Category Comments
Positive	Love	56,631	534,125	53,177	"সুন্দর"(nice), "ভালোবাসি"(love)	149,809
	Enthusiasm	37,965	523,566	64,533	"আকুল"(eager), প্লিজ(please)	
	Happy	26,596	200,403	24,929	"ভালো(good), "মাশাআল্লাহ"(masha Allah)	
	Fun	24,351	241,592	30,447	"হাসতে"(laugh), "মজা"(recreation)	
	Surprise	3501	16,715	4373	"চমৎকার"(excellent), "অবাক"(surprised)	
	Relief	765	3109	168	"জিতসো"(win), "বাচলাম"(survived)	
Negative	Angry	25,054	338,737	47,497	"নাস্তিক"(atheist), "শয়তান"(devil)	47,726
	Sad	6101	67,749	15,029	"আহারে"(scream), "কষ্ট"(trouble)	
	Sexual	5854	37,011	7173	"হট"(hot), "মাল"(sexy)	
	Disgust	3888	26,444	7177	"ফালতু"(nonsense), "আবাল"(stupid)	
	Boring	2494	14,925	2152	"ভাল্লাগেনা"(disliked), "চোরের"(thief's)	
	Worry	1810	10,693	541	"অসম্ভব"(impossible), "হবেনা"(never)	
	Fear	1442	7335	760	"মাফ"(forgive), "পারবা"(capable)	
	Hate	1083	4915	195	"ছি"(shit), "লজ্জা"(shame)	
Neutral	Neutral	3928	18,831	5266	"কেমন"(how), "আচ্ছা"(right)	3928
Total		**203,463**	**204,6150**	**164,974**		**203,463**

Table III
TRAINING SAMPLES GENERATION PROCESS FOR WORD EMBEDDINGS USING WINDOW SIZE 5

Source Bangla Text	Training Samples Generated From Source Text
আমি অনেক আগে থেকেই তোমাকে পছন্দ করি	(আগে, আমি) (আগে, অনেক) (আগে, থেকেই) (আগে, তোমাকে)
আমি অনেক আগে থেকেই তোমাকে পছন্দ করি	(থেকেই, অনেক) (থেকেই, আগে) (থেকেই, তোমাকে) (থেকেই, পছন্দ)
আমি অনেক আগে থেকেই তোমাকে পছন্দ করি	(তোমাকে, আগে) (তোমাকে, থেকেই) (তোমাকে, পছন্দ) (তোমাকে, করি)

Table IV
CHARACTER N-GRAM GENERATION PROCESS OF FASTTEXT MODEL

Word	Character Analysis	N-gram Length	Character N-grams
ভালোবাসা (love)	ভ +(া)আ +ল +(ে া)ও +ব +(া)আ +স+(া)আ	3	ভা, ভাল, আলো, লোব, ওবা, বাস, আসা, সা
ভালোবাসা (love)	ভ +(া)আ +ল +(ে া)ও +ব +(া)আ +স+(া)আ	4	ভাল, ভালো, আলোব, লোবা, ওবাস, বাসা, আসা

C. Data Preprocessing

The raw comments always contain irrelevant characteristics and noise, it is very important to eliminate them from the dataset [9].

1) Tokenization: During the tokenization process comments are divided into sentences and the sentences are divided into words. The Bangla comment "আপনার কাজগুলো আমার খুব ভালোলাগে", [i like your works very much] after tokenizing become "আপনার"(your), "কাজগুলো"(works), "আমার"(i), "খুব"(very), "ভালোলাগে"(like).

2) Punctuation, Non-Bangla Words and Emoticon Removal: The commonly used punctuation marks in Bangla are " ।", "?", "!", "-", "," etc and so on. Punctuation marks and special characters and symbols especially "#"(hashtags), @, & and braces have been excluded from the dataset. Non-Bangla words especially English words and unnecessary emoticons are also removed from one version of the dataset.

3) Stop Words Removal: The function words that are used repeatedly in a language but do not have any domain based meaning called stop words [10]. There are many noteworthy stop words available in Bangla such as "বরং"(rather), "কিন্তু"(but), "নতুবা"(or), "যদি"(if), "তুমি"(you), "অতএব"(therefore), "এখন"(now) etc. We have excluded all

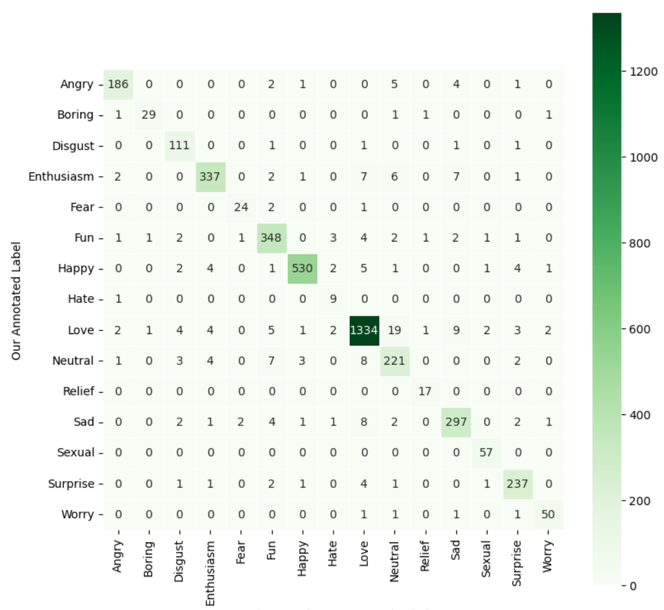

Figure 2. Confusion Matrix of Annotation Process

stop words from our developed dataset using the Bangla stop words[1] list.

4) Stemming: Stemming is the process of reducing a word to its base or root form [10]. For example, the words “চুক্তিটি”, “চুক্তিটিকে”, “চুক্তিটিতে”, "চুক্তিটির", "চুক্তিটা", "চুক্তিজনিত", "চুক্তিগুলো", "চুক্তিভিত্তিক" all have the same root word “চুক্তি”. By stemming these words, we can match them with other words that have the same root word, such as “চুক্তিগত” or “চুক্তিকৃত”. The Bangla comment "গতকাল মা ছাড়া শিশুটির ছবি ফেসবুকে দেখে ভীষন কষ্ট পেয়েছিলাম" become "গতকাল মা ছাড়া শিশু ছবি ফেসবুক দেখ ভীষন কষ্ট পেয়ে" after stemming is performed.

Table V
MAXIMUM ACCURACY OVERVIEW OF IMPLEMENTED ALGORITHMS

Domain	**Algorithm**	**Accuracy Without SMOTE (%)**	**Accuracy With SMOTE (%)**
ML	BNB	75.26	83.31
	DT	76.89	83.21
	LR	78.34	88.74
	KNN	69.95	81.58
	SVM	79.58	92.37
EL	RF	80.13	91.16
	XGB	79.92	92.55
	GB	80.03	91.87
DL	RNN	82.17	93.14
	LSTM	83.23	94.37
	Bi-LSTM	82.65	94.68
	CNN	82.91	94.55
	CNN-BiLSTM	**83.64**	**95.71**

D. Feature Extraction

Extracting relevant features from textual data is an important aspect in NLP [14]. Convert the raw textual data into some kind of strategic numerical form i.e. by extracting relevant features from texts broadly known as feature extraction.

1) Bag of Words: BOW is frequently used in NLP that works on the basis of term frequencies [1-2]. Considering a Bangla sample corpus of three comments as "আমার ভালো লাগছে না।", "আমার মন খারাপ।", "আমার সাথে দুষ্টামি করো না।", [i do not feel good. i am upset. do not mess with me], counting BOW and get the features as "আমার"(my): 3, "ভালো"(good): 1, "লাগছে"(feel): 1, "না"(no): 2, "মন"(am): 1, "খারাপ" (upset): 1, "সাথে"(with): 1, "দুষ্টামি"(mess): 1, "করো"(do): 1.

2) TF-IDF: It is the most widely used feature extraction metric for NLP based classification tasks [5] which is calculated according to the equation (3).

$$TF = \frac{\text{Frequency of a word in a comment}}{\text{Total no. of words in the comment}} \tag{1}$$

$$IDF = \log_e \frac{\text{Total number of comments}}{\text{No. of comments contain the word}} \tag{2}$$

$$TF - IDF = TF \times IDF \tag{3}$$

[1] https://www.ranks.nl/stopwords/bengali

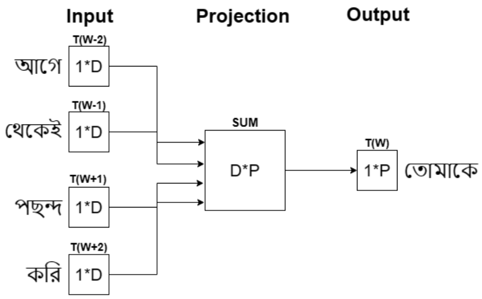

Figure 3. Word Embedding Using CBOW

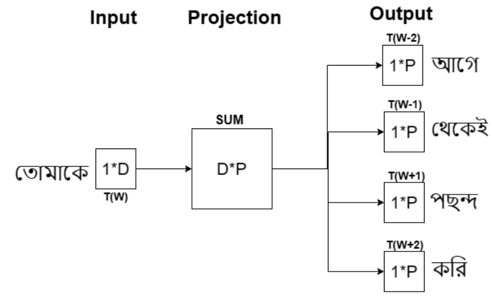

Figure 4. Word Embedding Using Skipgram

3) TF-IDF-ICF: The term ICF stands for Inverse Class Frequency introduced by Wang et al. [14] which is calculated according to the equation (4):

$$ICF = \log_e 1 + \frac{\text{No. of total categories}}{\text{No. of categories contain the word}} \tag{4}$$

$$TF - IDF - ICF = TF \times IDF \times ICF \tag{5}$$

4) Word Embeddings: Word Embedding is a vector based feature extraction metric used in NLP [15]. It can be implemented using Word2Vec, FastTest or GloVe methods based on the mechanism CBOW or skipgram [15]. Table III depicts the generation process of the training samples for word embeddings. The CBOW model learns through context words and tries to predict the target word (Figure 3) whereas skipgram model tries to predict its neighbors (context words) using the current word (Figure 4). An extension of Word2vec model that works on the basis of character level n-grams (Figure IV), it is known as FastText model which is good for morphological rich language like Bangla [15]. Another word embedding model called GloVe (Global Vector) works based on global corpus statistics and create fixed sized vectors using a co-occurrence matrix [6]. Bangla-BERT is another pre-trained model developed by Google AI works on the basis of transformers and attention mechanism [6].

E. Bangla Sentiment Analysis Model

To observe the significance of different hybrid feature extraction techniques on Bangla SA, we have implemented

Table VI

Obtained Results of Applied Algorithms Using Different Feature Extraction Techniques Based on 15 Sentiment Categories

Feature	**Accuracy (%)**												
	BNB	**DT**	**LR**	**KNN**	**SVM**	**RF**	**XGB**	**GB**	**RNN**	**LSTM**	**Bi-LSTM**	**CNN**	**CNN-BiLSTM**
BOW+2-Gram	61.13	73.25	74.51	39.78	77.63	80.31	79.88	80.03	N/A	N/A	N/A	N/A	N/A
BOW+3-Gram	61.33	73.29	74.50	39.91	77.67	80.23	79.87	79.98	N/A	N/A	N/A	N/A	N/A
TF-IDF+2-Gram	65.72	78.93	77.67	42.96	80.34	82.52	82.31	82.19	N/A	N/A	N/A	N/A	N/A
TF-IDF+3-Gram	65.81	78.97	77.98	43.10	80.77	82.63	81.69	82.33	N/A	N/A	N/A	N/A	N/A
TF-IDF-ICF+2-Gram	64.33	80.14	78.40	43.22	82.19	83.12	82.13	82.44	N/A	N/A	N/A	N/A	N/A
TF-IDF-ICF+3-Gram	64.27	80.31	78.42	43.23	82.76	83.12	82.15	82.53	N/A	N/A	N/A	N/A	N/A
Word2Vec+CBOW (gensim)	28.71	58.16	60.67	57.90	62.61	71.18	70.13	70.69	72.39	73.65	74.68	75.89	82.31
Word2Vec+Skipgram (gensim)	35.95	57.84	62.94	60.62	67.98	71.79	71.05	69.98	74.23	75.63	76.92	80.39	83.26
Word2Vec+CBOW+ Skipgram (gensim)	28.55	57.81	60.43	56.89	64.69	71.19	69.72	68.87	73.13	76.98	75.97	80.14	82.79
Word2Vec+CBOW (tensorflow)	24.13	55.79	51.07	43.26	60.53	64.67	63.19	62.74	79.54	81.26	82.39	81.37	83.76
Word2Vec+Skipgram (tensorflow)	25.26	57.81	54.32	44.17	61.29	64.98	63.55	61.87	79.98	82.05	82.31	82.91	82.77
Word2Vec+CBOW+ Skipgram (tensorflow)	24.89	55.86	53.68	43.29	60.95	64.83	63.14	61.68	78.81	82.39	80.39	79.68	81.26
FastText+CBOW	56.82	67.64	71.62	**72.69**	72.86	76.21	73.42	72.87	77.37	79.81	80.45	80.11	82.38
FastText+Skipgram	**66.85**	67.78	71.61	72.58	72.59	76.29	74.51	73.27	79.91	81.29	84.25	83.22	84.27
FastText+CBOW+ Skipgram	59.87	67.74	71.64	72.62	72.89	76.23	75.85	72.96	78.63	80.94	82.71	81.21	83.57
GloVe+CBOW	55.72	66.52	71.23	57.89	72.43	74.89	75.76	70.83	76.39	78.35	79.42	78.98	80.53
GloVe+Skipgram	65.35	67.39	70.54	58.93	71.99	75.61	74.38	69.87	78.13	75.53	81.32	80.32	81.26
GloVe+CBOW+ Skipgram	61.37	65.74	71.27	64.79	70.99	72.94	71.29	70.89	73.41	78.91	79.47	78.96	80.93
Bangla-BERT+CBOW	63.47	84.26	83.49	67.91	84.67	85.79	83.41	84.62	88.89	86.41	86.49	87.91	89.13
Bangla-BERT+Skipgram	64.36	83.19	**83.77**	66.43	**84.88**	85.34	**83.59**	**84.91**	**88.91**	85.73	**85.96**	**88.33**	**90.24**
Bangla-BERT+CBOW+ Skipgram	64.53	**84.37**	83.73	67.05	84.79	**85.77**	82.97	84.77	88.86	**86.90**	86.81	88.13	89.91

Table VII

Obtained Results of Applied Algorithms Using Different Feature Extraction Techniques Based on 3 Sentiment Categories

Feature	**Accuracy (%)**												
	BNB	**DT**	**LR**	**KNN**	**SVM**	**RF**	**XGB**	**GB**	**RNN**	**LSTM**	**Bi-LSTM**	**CNN**	**CNN-BiLSTM**
BOW+2-Gram	80.32	82.37	83.04	57.33	82.43	83.78	82.47	82.88	N/A	N/A	N/A	N/A	N/A
BOW+3-Gram	80.13	81.69	82.34	59.75	81.66	83.96	82.39	81.89	N/A	N/A	N/A	N/A	N/A
TF-IDF+2-Gram	82.69	88.23	88.04	58.39	89.47	90.33	90.45	89.97	N/A	N/A	N/A	N/A	N/A
TF-IDF+3-Gram	83.05	88.21	88.19	61.46	89.48	90.41	90.23	90.17	N/A	N/A	N/A	N/A	N/A
TF-IDF-ICF+2-Gram	83.18	88.95	88.61	58.20	89.92	91.07	90.86	89.99	N/A	N/A	N/A	N/A	N/A
TF-IDF-ICF+3-Gram	**83.31**	89.02	**88.74**	59.16	90.01	**91.16**	89.56	90.45	N/A	N/A	N/A	N/A	N/A
Word2Vec+CBOW (gensim)	65.82	78.83	80.84	81.11	81.86	85.78	84.35	83.26	83.49	83.99	84.74	84.56	89.57
Word2Vec+Skipgram (gensim)	69.19	79.92	81.46	80.31	83.20	86.44	84.53	84.36	82.43	87.94	86.24	85.41	90.13
Word2Vec+CBOW+ Skipgram (gensim)	64.83	78.54	80.91	81.09	81.89	85.99	84.49	85.23	84.58	86.78	89.05	88.17	92.48
Word2Vec+CBOW (tensorflow)	54.72	73.35	75.66	71.26	78.59	80.46	82.33	81.16	83.45	85.90	87.43	89.99	89.93
Word2Vec+Skipgram (tensorflow)	56.13	77.91	78.52	74.89	79.94	80.09	84.06	83.34	84.65	86.19	89.72	87.70	92.30
Word2Vec+CBOW+ Skipgram (tensorflow)	54.97	77.18	76.99	74.57	78.92	79.99	84.14	82.73	81.94	86.31	90.18	90.02	91.61
FastText+CBOW	65.82	76.64	80.61	81.11	81.86	85.21	86.49	84.45	85.73	89.90	88.15	91.15	92.35
FastText+Skipgram	75.85	76.78	81.27	**81.58**	82.34	85.29	87.95	88.21	89.55	90.31	92.19	90.95	93.54
FastText+CBOW+ Skipgram	75.83	76.63	80.97	81.61	81.93	85.23	85.69	87.89	90.14	90.78	91.44	91.08	93.25
GloVe+CBOW	67.23	75.19	78.47	79.63	80.96	85.35	84.14	86.51	88.90	86.05	87.68	89.96	91.55
GloVe+Skipgram	73.89	75.82	80.94	80.38	81.53	85.67	85.19	88.98	84.93	87.34	89.95	90.42	92.33
GloVe+CBOW+ Skipgram	76.12	76.07	79.57	80.44	81.29	85.12	85.32	87.93	85.66	86.26	89.92	90.49	92.28
Bangla-BERT+CBOW	80.13	83.16	86.91	78.92	92.14	91.03	92.34	90.68	92.89	94.13	94.45	93.44	94.47
Bangla-BERT+Skipgram	82.15	83.17	87.72	80.14	**92.37**	91.09	**92.55**	**91.87**	93.03	**94.37**	94.66	94.26	**95.71**
Bangla-BERT+CBOW+ Skipgram	81.99	**83.21**	87.79	79.54	92.29	91.14	92.47	91.74	**93.14**	94.26	**94.68**	**94.55**	95.27

several ML, EL and DL algorithms. We have implemented the Synthetic Minority Oversampling Technique (SMOTE) to balance our dataset and TABLE V describes the effects of it on the performance of the applied methods.

IV. Experimental Results Analysis And Discussion

The proposed work mainly focuses on discovering an efficient feature extraction metric for Bangla SA. In this

Table VIII
Comparison Among Existing Works and Our Proposed Work

Name	Year	Dataset Used	No. of Categories	Feature Metric	Used Model	F1-Score	Accuracy (%)
N. R. Bhowmik et al. [4]	2022	2900 & 2600	3	Word2Vec	BERT-LSTM	N/A	84.18
N. J. Prottasha et al. [6]	2022	2900 & 2600 & 4 others	3	BERT	CNN-BiLSTM	0.93	94.15
M. Kabir et al. [1]	2023	158,065	3	BOW+ N-gram	BERT	0.93	N/A
A. K. Bitto et al. [3]	2023	1400	2	Word2Vec	LSTM	0.85	91.07
Our Proposed	**2023**	**203,463**	**3**	**BERT+ Skipgram**	**CNN-BiLSTM**	**0.96**	**95.71**
			15	**BERT+ Skipgram**	**CNN-BiLSTM**	**0.91**	**90.24**

Table IX
Best Performance Measurements of Different Domains Using (Bangla-BERT+Skipgram) Based on 3 categories

Domain	Best Model	Precision	Recall	F1-Score	Accuracy (%)
ML	SVM	0.96	0.91	0.93	92.37
EL	XGB	0.89	**0.99**	0.94	92.55
DL	CNN-BiLSTM	**0.97**	0.94	**0.95**	**95.71**

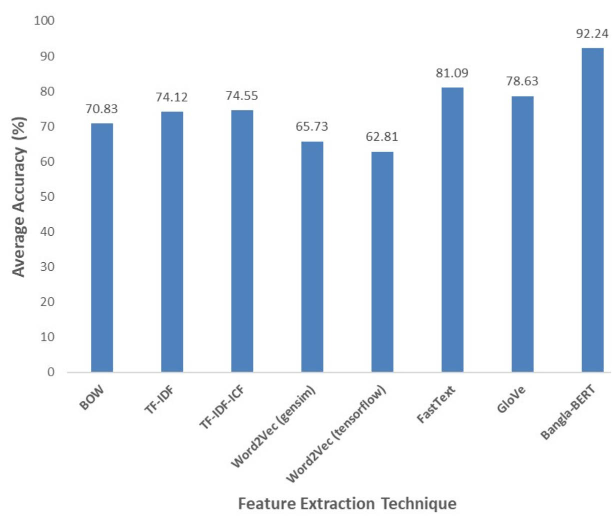

Figure 5. Average Performance of Different Feature Extraction Techniques

work, we have developed total 21 different hybrid feature extraction techniques and implemented ML, EL and DL algorithms to evaluate them. The obtained results of applied algorithms based on 15 and 3 categories are shown in TABLES VI and VII respectively. The best achieved results are highlighted in bold signs. Among ML algorithms, SVM (accuracy 84.88%) outperforms all other methods using (Bangla-BERT+Skipgram) while EL and DL algorithms achieved highest accuracy 85.77% (Bangla-BERT+CBOW+Skipgram) and 90.24% (Bangla-BERT+Skipgram) using RF and the hybrid model CNN-BiLSTM. The obtained results for 3 categories are summarized ans shown in TABLE IX. The TF-IDF-ICF also performed well and obtained better accuracy than TF-IDF. The experimental results show that Skipgram outperforms CBOW (observe TABLES VI and VII). The comparison among existing works and our proposed work is illustrated in TABLE VIII where we have measured the dataset used, no. of categories, feature metric, used model, f1-score and accuracy as comparative features. We have examined SA on both 15 categories and 3 categories and noticed that accuracy (i.e. results) and no. of categories are proportional to each other. The best achieved accuracy is 90.24% in 15 categories and 95.71% in 3 categories. The average performance of different feature extraction methods is summarized in Figure 5, where Bangla-BERT outperms all other methods with an average accuracy of 92.24%.

V. Conclusion

Sentiment Analysis (SA) or opinion mining is a topic of great importance in every language. But it is a matter of regret that a few research works have been conducted in Bangla language on SA compared to other languages due to lack of resources. This work focuses on examining different hybrid feature extraction techniques on Bangla SA using a new comprehensive dataset with ML, EL, and DL approaches. The proposed dataset contain 203,493 Bangla comments collected from 5 microblogging sites. This work examined 21 different hybrid feature extraction techniques and found the novel method (Bangla-BERT+Skipgram) which outperforms all other techniques. Among the word embedding models FastText performs better than Word2Vec and GloVe. TF-IDF-ICF is also explored in this study, which outperforms TF-IDF for ML approaches. Average accuracy achieved for the proposed Bangla-BERT feature extraction method is 92.24%. In the future, we want to enrich and balance our dataset more and convert it to a representative one for Bangla SA and explore evolutionary algorithms for extracting features from texts.

References

[1] M. Kabir, O. B. Mahfuz, S. R. Raiyan, H. Mahmud, M. K. Hasan, “BanglaBook: A Large-scale Bangla Dataset for Sentiment Analysis from Book Reviews,” Computation and Language, 2023, https://doi.org/10.48550/arXiv.2305.06595.

[2] N. Nafisa, S. J. Maisha, A. K. M. Masum, “Document Level Comparative Sentiment Analysis of Bangla News Using Deep Learning-based Approach LSTM and Machine Learning Approaches,” Applied Intelligence for Industry 4.0, pp. 198-211, 2023.

[3] A. K. Bitto, M. H. I. Bijoy, M. S. Arman, I. Mahmud, A. Das, J. Majumder, “Sentiment Analysis from Bangladeshi Food Delivery Startup Based on User Reviews Using Machine learning and Deep Learning,” Bulletin of Electrical Engineering and Informatics, vol. 12, no. 4, pp. 2282-2291, 2023.

[4] N. R. Bhowmik, M. Arifuzzaman, M. R. H. Mondal, “Sentiment Analysis on Bangla Text Using Extended Lexicon Dictionary and Deep Learning Algorithms,” Array, vol. 3, no. 2590-0056, pp. 100123, 2022, https://doi.org/10.1016/j.array.2021.100

[5] M. Hassan, S. Shakil, N. N. Moon, M. M. Islam, R. A. Hossain, A. Mariam, F. N. Nur, “Sentiment Analysis on Bangla Conversation Using Machine Learning Approach,” International Journal of Electrical and Computer Engineering (IJECE), vol. 12, no. 5, pp. 5562-5572, 2022, 10.11591/ijece.v12i5.pp5562-5572.

[6] N. J. Prottasha, A. A. Sami, M. Kowsher, S. A. Murad, A. K. Bairagi, M. Masud, M. Baz, “Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning,” Sensors, vol. 22, no. 11, pp. 4157, 2022.

[7] N. Alvi, K. H. Talukder, A. H. Uddin, “Sentiment Analysis of Bangla Text Using Gated Recurrent Neural Network,” International Conference on Innovative Computing and Communications Advances in Intelligent Systems and Computing, vol 1388, 2022, https://doi.org/10.1007/978-981-16-2597-8_7

[8] M. A. Shafin, M. M. Hasan, M. R. Alam, M. A. Mithu, A. U. Nur, M. O. Faruk, “Product Review Sentiment Analysis by Using NLP and Machine Learning in Bangla Language,” 2020 23rd International Conference on Computer and Information Technology (ICCIT), pp. 1-5, 2020, doi: 10.1109/ICCIT51783.2020.9392733.

[9] R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter, A. K. Das, “An Automated System of Sentiment Analysis from Bangla Text Using Supervised Learning Techniques,” 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), pp. 360-364, 2019, doi: 10.1109/CCOMS.2019.8821658.

[10] A. Akther, M. S. Islam, H. Sultana, A. R. Rahman, S. Saha, K. M. Alam, R. Debnath, “Compilation, Analysis and Application of a Comprehensive Bangla Corpus KUMono,” IEEE Access, vol. 10, pp. 79999-80014, 2022, doi: 10.1109/ACCESS.2022.3195236.

[11] M. A. Rahman, E. k. Dey, “Datasets for Aspect-Based Sentiment Analysis in Bangla and Its Baseline Evaluation,” Data 3, no. 2, pp. 15, 2018, https://doi.org/10.3390/data3020015.

[12] M. H. Alam, M. M. Rahoman, M. A. K. Azad, “Sentiment Analysis for Bangla Sentences Using Convolutional Neural Network,” 2017 20th International Conference of Computer and Information Technology (ICCIT), pp. 1-6, 2017, doi: 10.1109/ICCITECHN.2017.8281840.

[13] A. Hassan, M. R. Amin, A. K. A. Azad, N. Mohammed, “Sentiment Analysis on Bangla and Romanized Bangla Text Using Deep Recurrent Models,” 2016 International Workshop on Computational Intelligence (IWCI), pp. 51-56, 2016, doi: 10.1109/IWCI.2016.7860338.

[14] D. Wang, H. Zhang, “Inverse-Category-Frequency based Supervised Term Weighting Scheme for Text Categorization,” arXiv preprint arXiv:1012.2609, 2010.

[15] S. H. Sumit, M. Z. Hossan, T. Al Muntasir, T. Sourov, “Exploring Word Embedding for Bangla Sentiment Analysis,” 2018 International Conference on Bangla Speech and Language Processing, pp. 1-5, 2018, 10.1109/ICBSLP.2018.8554443.