Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 11:2023
DOI: 10.5281/zenodo.10152990

SENTIMENT ANALYSIS OF BANGLA-ENGLISH CODE-MIXED AND TRANSLITERATED SOCIAL MEDIA COMMENTS USING MACHINE LEARNING

FAHIMA HOSSAIN
Lecturer, Department of Computer Science & Engineering, Hamdard University Bangladesh, Bangladesh. E-mail: minda.fahima25@gmail.com

NUSRAT JAHAN
BSc in CSE, Hamdard University Bangladesh, Bangladesh. E-mail: nusrat.jahan.az@gmail.com

JOYREYA ABADIN
BSc in CSE, Hamdard University Bangladesh, Bangladesh. E-mail: ajoyreya285@gmail.com

Abstract

In this era of technology, online communication has expanded to incorporate different language combinations and expressive styles. Code-mixing, or combining languages such as Bengali and English, is frequent in multilingual environments. In addition, the use of emojis, which are small graphical icons, has increased the complexity of online communication by allowing for more emotional expression. Sentiment analysis of comments on social media that are combined with emojis and written in Bangla-English language is the main topic of this work. This research aims to understand and classify the sentiments expressed in code-mixed comments, including the nuanced role of emojis. The data was preprocessed first, then feature extraction was performed using the TF-IDF Vectorizer and CountVectorizer algorithms. For the analysis, nine different machine learning algorithms were used. With a remarkable accuracy of 85.7% and an F1 score of 85.0%, the Support Vector Classifier stood out as the most successful model. This highlights the utility of including emoji-based features for complex sentiment analysis, particularly when dealing with code-mixed data. The dataset contained 2055 comments from Facebook pages, including Bangla-English comments with and without emojis and comments that only contained emojis. To prepare the dataset for analysis, preprocessing methods included removing irrelevant data and converting emojis into Unicode short names.

Index Terms: Sentiment Analysis, Emoji Analysis, Natural Language Processing, Machine Learning, Code-mixing, Emoji, Unicode.

1 INTRODUCTION

Today, individuals effortlessly blend languages while communicating online, which is referred to as code-mixing, in order to express their thoughts and emotions. People who are fluent in multiple languages often blend them together, incorporating English elements into their native language. The fusion of English and Bengali, known as Benglish, is a prime example of this phenomenon. It enables individuals to convey their thoughts and ideas by using a blend of both languages. This linguistic phenomenon highlights the ever-changing nature of communication within multilingual groups [1].

Sentiment analysis, a crucial area of natural language processing (NLP), focuses on understanding and classifying the emotions and points of view expressed in textual data. According to the analytic context, these attitudes are divided into a variety of groups, including negative, neutral, and positive, with the possibility to further segment them into extremely negative, strongly positive, and other categories [2]. This technique has a wide range of applications, from business insights to election forecasts, by decoding feelings in customer feedback, social media posts, and public opinion [3, 4]. However, whereas sentiment analysis research thrives in English, Bangla lags behind due to linguistic difficulty and a lack of resources [5].

Emojis have revolutionized modern communication and have become a common feature in social media conversations [6]. Through the effective communication of emotions, context, and intent, these Unicode symbols cross linguistic and cultural boundaries [7, 8, 9]. On Messenger, more than 900 million emoticons are exchanged every day, while Facebook sees more than 700 million emojis shared in a day [10]. Emojis can substitute for words and represent a variety of ideas, including emotions, objects, activities, and more because of their adaptability [11]. Emoji usage is widespread, but little study has been done on it [12, 13, 14]. Emojis and text are combined in this study, which recognizes the variety of expressions they provide. Including emojis in textual sentiment analysis is beneficial. They provide a universal language [7, 8, 9] for expressing emotions, enhance prediction accuracy, and align with contemporary communication styles.

This study collects Facebook data, which consists of Bangla-English comments with and without emojis, along with comments that only contain emojis. The dataset contains a total of 2055 labeled points and has been preprocessed and balanced using TF-IDF and Count Vectorizer before being split into training and testing. To perform sentiment analysis, nine categorization algorithms are utilized.

This work aims to fill the research gap on extracting sentiment from code-mixed Bengali-English texts with emojis in social media. The contributions of this study are as follows:

1. Created a dataset of tagged Bangla-English code-mix with added emojis.
2. Managed imbalanced data properly to reduce overfitting.
3. Created a customized Bangla-English lemmatizer.
4. Developed an empirical model for sentiment analysis.

This research is divided into numerous Sections that methodically cover various parts of the investigation. Section 2 offers a concise summary of related works to determine the current state of knowledge in the subject matter. In Section 3, a summary of the data-gathering process is provided, including details on data storage and the methodology used for the study. Furthermore, Section 4 evaluates the proposed model and presents its outcomes, providing significant information on its effectiveness. The findings of the paper are presented in Section 5, which also suggests possible directions for further research.

2. LITERATURE REVIEW

Mahmud AA et al. [15], proposed an approach to sentiment analysis that supports Bangla-English code-mixing and transliterated features. The authors pre-processed the data by stopword remove, punctuation remove, pos tagging and for word embedding they used glove, and word2vector, and after all that, LSTM and BERT were applied.

Khan et al. [16], proposed sentiment analysis approach for Bangla language and worked with five different emotions: Happy, Sad, Angry, Surprised, and Excited. Additionally, their paper also deals with two categories: Abusive and Religious. The authors used the TFiDF method for converting the data into weighted vector form. The total work has done in two experimental steps. In the first experiment, they trained the dataset with algorithms. In the second experiment, they mapped their 7 classes (Sad, Happy, Angry, Excited, Religious, Abusive, Surprised) into three higher classes (Positive, Negative, and Neutral). The best accuracy (62%) for 7 class achieved using Support Vector Machine (SVM). And for 3 classes, 73% is the best accuracy obtained using Support Vector Machine (SVM).

Khan et al [17], proposed a model for depression analysis from Bangla post via Natural Language Processing (NLP). For data cleaning they added contraction, removed regular expression, removed stop words and tokenized the data using CountVectorizer. The Multinomial Naive Bayes algorithm provide best accuracy for their model.

Al-Azani et al. [2], introduced a approach by incorporating nonverbal features, specifically emojis, for sentiment analysis of Arabic microblogs where each instances, contained at least one emoji. For feature selection, they used ReliefF and Correlation-Attribution Evaluator (CAE) as their approach was text-independent and would only focus on emojis features. The study achieves an F1 score of 80.30% and an AUC (Area Under the Curve) of 87.30% by selecting multinomial naive Bayes classifier and 250 of the most relevant emojis. The authors indicate that using emoji-based features alone can be highly effective in detecting sentiment polarity.

Mandal et al. [18], addressed sentiment analysis for code-mixed social media content. They created a Bengali-English dataset and utilized hybrid approaches for language and sentiment tagging, obtaining 81% accuracy in language identification and 80.97% accuracy in sentiment classification. The research paper provides valuable resources for code-mixed sentiment analysis, which could potentially inspire more research in this complex field of natural language processing.

Singh et al. [1], describe their findings on sentiment analysis of Hinglish (English-Hindi mix) code-mixed social media tweets. Data consolidation, cleansing, transformation, and modeling are all part of the research. There were several vectorization approaches and machine learning algorithms utilized. The SemEval-2020 dataset was divided into positive, neutral, and negative feelings. The ensemble voting classifier attained the best F1-score (69.07), suggesting the possibility for additional improvements employing neural networks and advanced data processing techniques. The research underlines the need to normalize Hindi spellings and discusses potential improvements for Hindi words using neural networks and methods such as lemmatization and part-of-speech tagging.

Alzubaidi et al. [14], employed heterogeneous data mining (HDM) for integrated processing to explore sentiment analysis of tweets utilizing both text and emojis. The results indicate that using HDM along with text or emojis helps improve sentiment analysis accuracy. With Support Vector Machines (SVM), the accuracy reached 85.89%. This highlights the significance of incorporating emojis to enhance bi-sense sentiment analysis accuracy.

Tareq et al. [19], in this study the author's utilized a Bangla-English code-mixed (BE-CM) dataset they created to address challenges related to sentiment analysis in low-resource languages like Bangla. Their data augmentation technique improves cross-lingual comprehension in code-mixed expressions, achieving an 87% weighted F1 score using XGBoost and FastText embeddings. The work tackles the issues of sentiment analysis in multilingual communities, particularly for languages with limited resources like Bangla.

Jamatia et al. [20], the paper research sentiment analysis in Indian language code-mixed social media texts. For sentiment prediction in English-Hindi and English-Bengali, they compare standard machine learning approaches with deep learning models such as BiLSTM-CNN, Double BiLSTM, Attention-based, and BERT. The deep learning models consistently outperformed the classical methods, which is pretty impressive. The Attention-based model achieved the highest F1 scores of 67.5 (English-Bengali) and 60.4 (English-Hindi). It's interesting to note that while domain-specific BERT embeddings show potential, pre-trained BERT embeddings are still pretty competitive. Overall, the study highlighted the challenges in code-mixed sentiment analysis and emphasized the value of deep learning as well as the need for more extensive datasets.

Limitations: The existing literature reveals a variety of approaches to sentiment analysis, with some focusing solely on code-mix languages, others on emojis exclusively, and some on text and emojis in native languages. However, there is still a lack of research on Bangla-English code-mixed language, where English elements are integrated into native language text through phonetic typing. Our study presents a sentiment analysis framework for Bangla-English code-mixed language with emojis, as no previous exploration has been conducted on this topic yet. The studies discussed in this section are summarized in Table 1 based on various attributes.

Table 1: Overview of Relevant Studies Referenced from [1, 2 14-20].

Authors	Contributions	Datasets	Class	Performances (Best Score)
Mahmud AA et al. [15]	1. Applied two types of feature extraction methods to convert text into vectors. 2. Plotted 100 dimensions' vectors into 2 dimensional graphs. 3. Used test set differently to measure the performance of the model. The first test set contains 10% code mix data and the second	Bkash app review from google playstore (0.2 million sample with 12% bangla-english mixed sentence)	Positivie, Negative	Accuracy: 94%

	test set contains 42% code mix data.			
Khan et al. [16]	1. Created dataset of pure Bangla text or Bangla text mixed with some characters. 2. Applied Tf-idf as feature extraction	Facebook post comments	Positivie, Negative, Neutral	Accuracy: 73%
Khan et al [17]	1. Build a model to detect depression related post. 2. Applied Tf-idf as feature extraction	Bengali post of facebook	Happy, Sad	Accuracy: 86.67%
Al-Azani et al [2]	1. Conducted research on Arabic sentiment analysis to evaluate the impact of different emojis on predicting sentiment polarity.	Arabic microblogs (2091 sample)	Positivie, Negative	Accuracy: 80.34%
Soumil Mandal et al. [18]	1. Create a gold standard Bengali-English code-mixed dataset. 2. Develop an accurate language identification system with an 81% accuracy rate. 3. The dataset, which includes code-mixed text and is annotated with language and polarity tags, is now accessible in JSON format for future research purposes.	Gold standard Bengali-English code-mixed dataset (600 tweets)	Positive, Negative, And Neutral	Accuracy: 80.97%
Singh et al. [1]	1.Applied different data-cleaning approaches, transformation techniques, and machine-learning algorithms. 2. Identified areas for growth and development, such as neural network exploration and spell normalization.	Task 9: Sentiment Analysis of Code-Mixed Social Media Text (Hinglish) (14,000 tweets) [21]	Positive, Negative, And Neutral	F1-score: 69.07%
Alzubaidi et al. [14]	1. Uses texts and emojis to demonstrate HDM's supremacy for sentiment analysis. 2. Highlights the efficacy of emojis even in small quantities. 3. Introduces a uniform approach for both text and emoji processing.	English corpus composed of 16,207 tweets [22]	Positive and Negative	Accuracy: 85.89%
Tareq et al. [19]	1. Creating the BE-CM in code-mixed Bangla-English text. 2. Evaluating several models and BERT iterations for sentiment categorization. 3. Introducing an effective data augmentation strategy for	Gold standard Bangla-English code mix (BE-CM) dataset (18,074 Sentences)	Positive state, Negative state, Mixed positive, Mixed	F1- score: 87%

	cross-lingual sentiment analysis.		negative and Neutral	
Jamatia et al. [20]	1. Demonstrates the effectiveness of deep learning in code-mixed sentiment analysis. 2. Increse sentiment analysis in multilingual contexts	ICON-2017 SAIL Tool-Contest Annotated Corpora [43], Joshi et al. [24], SemEval 2013 Task 2B Twitter Dataset [25], Barnes et al. [26]	Positive, Negative, And Neutral	F1-scores: 67.5% (English-Bengali) and 60.4% (English-Hindi)

3. METHODOLOGY

The proposed model working mechanism to detect sentiment is present in this section. Figure 1 illustrates the approach used in this model.

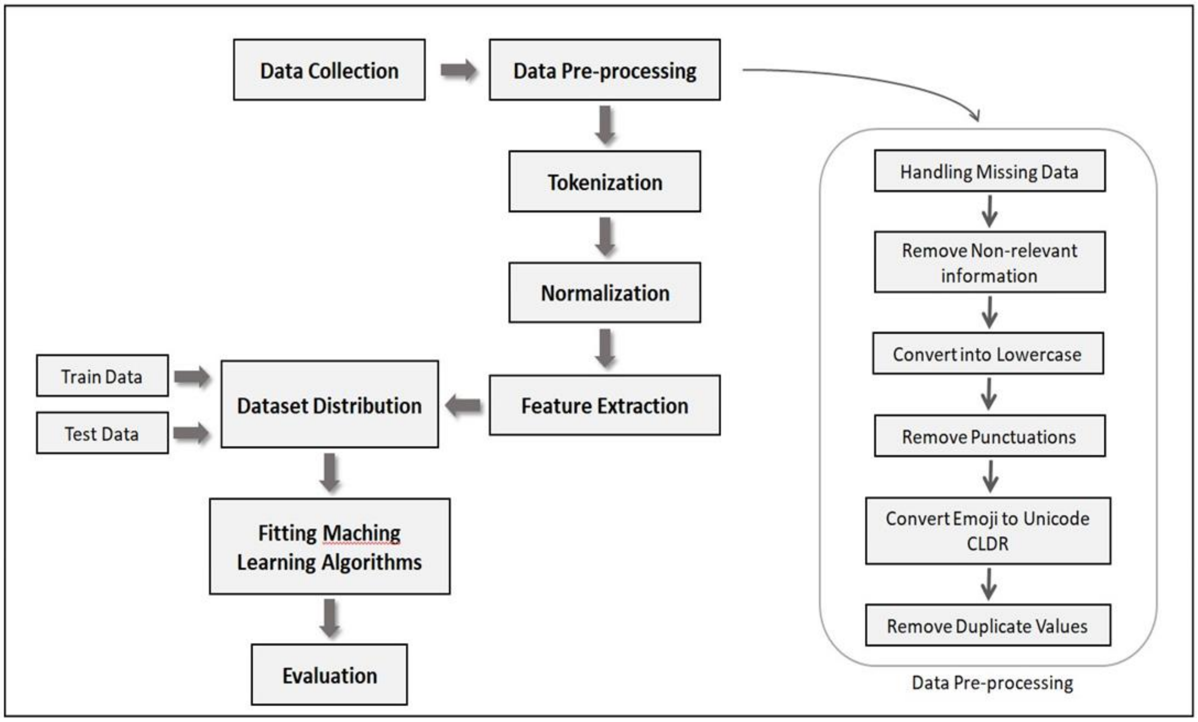


Fig. 1: Flow Diagram of the Proposed Model.

3.1 Data collection and dataset preparation

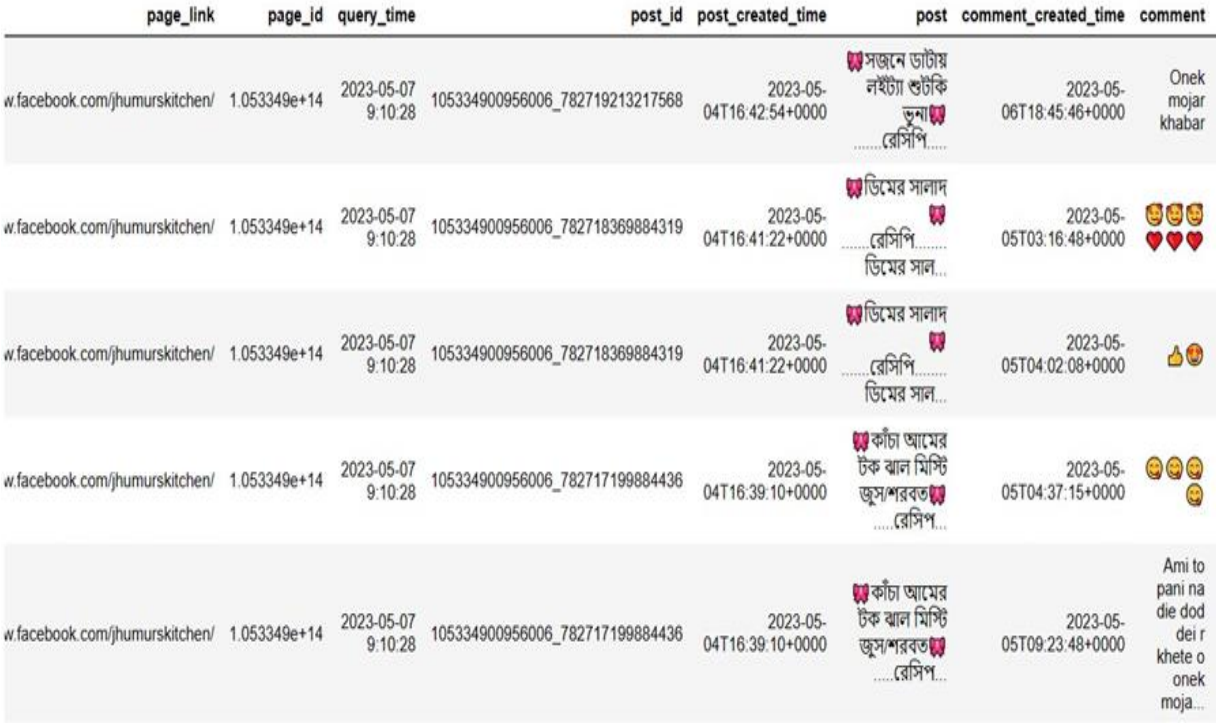
Machine learning heavily relies on data as a fundamental component. The data used to utilize this paper is the comments of general people that have been collected manually from different Facebook pages using facepager. Facepager uses a JSON-based API to fetch publicly accessible data from Facebook, YouTube, and Twitter. An access token is required, generated automatically when a profile is logged in. An additional method for obtaining access tokens is through the utilization of a website called the 'Graph API

Explorer - Facebook Developer.' Facepager exports raw data as an Excel CSV sheet and stores it in a SQLite database (2).

2055 comments have been collected that are mainly Bangla, written using English letters (generally known as Banglish), mix of Banglish comments with or without emoji, or only emoji. Along with these, some commonly used English words and sentences that are mostly used (such as: ok, well done, good, problem, best of luck, etc) are also consumed as collected comment. The collected comments are then exported as an Excel CSV sheet. Each comment has been labeled manually based on its real aspect of use. Overall nine columns are present in the dataset that is explained in Table 2. Figure 1 represents the sample scenario of the dataset.

Table 2: Columns included in the dataset.

Column	Description
Page_link	The page link of particular post from which comments are collected
Page_id	The page id of particular post from which comments are collected
Query_time	The time when the comments have collected
Post_id	The post id of post from which comments are collected
Post_created_time	The post created time of post from which comments are collected
Post	The post from which comments are collected
Comment_created_time	Commented time
Comment	Comment of general people
Sentiment	Class (positive, Negative, Neutral) of comment



page_link	page_id	query_time	post_id	post_created_time	post	comment_created_time	comment
w.facebook.com/jhumurskitchen/	1.053349e+14	2023-05-07 9:10:28	105334900956006_782719213217568	2023-05-04T16:42:54+0000	🎀সজনে ডাটায় লইট্যা শুটকি ভুনা🎀 ....রেসিপি.....	2023-05-06T18:45:46+0000	Onek mojar khabar
w.facebook.com/jhumurskitchen/	1.053349e+14	2023-05-07 9:10:28	105334900956006_782718369884319	2023-05-04T16:41:22+0000	🎀ডিমের সালাদ 🎀 ....রেসিপি........ ডিমের সাল...	2023-05-05T03:16:48+0000	🥰🥰🥰❤️❤️❤️
w.facebook.com/jhumurskitchen/	1.053349e+14	2023-05-07 9:10:28	105334900956006_782718369884319	2023-05-04T16:41:22+0000	🎀ডিমের সালাদ 🎀 ....রেসিপি........ ডিমের সাল...	2023-05-05T04:02:08+0000	👍😍
w.facebook.com/jhumurskitchen/	1.053349e+14	2023-05-07 9:10:28	105334900956006_782717199884436	2023-05-04T16:39:10+0000	🎀কাঁচা আমের টক ঝাল মিস্টি জুস/শরবত🎀 .....রেসিপ...	2023-05-05T04:37:15+0000	😋😋😋😋
w.facebook.com/jhumurskitchen/	1.053349e+14	2023-05-07 9:10:28	105334900956006_782717199884436	2023-05-04T16:39:10+0000	🎀কাঁচা আমের টক ঝাল মিস্টি জুস/শরবত🎀 .....রেসিপ...	2023-05-05T09:23:48+0000	Ami to pani na die dod dei r khete o onek moja...

Fig. 2: The sample scenario of the dataset.

3.2 Data preprocessing

To enhance the performance of the machine learning model raw data needs to be preprocessed. At the very first step of data preprocessing, keeping only comments and sentiment columns, all other unnecessary columns have been dropped. The preprocessing steps followed for this model are:

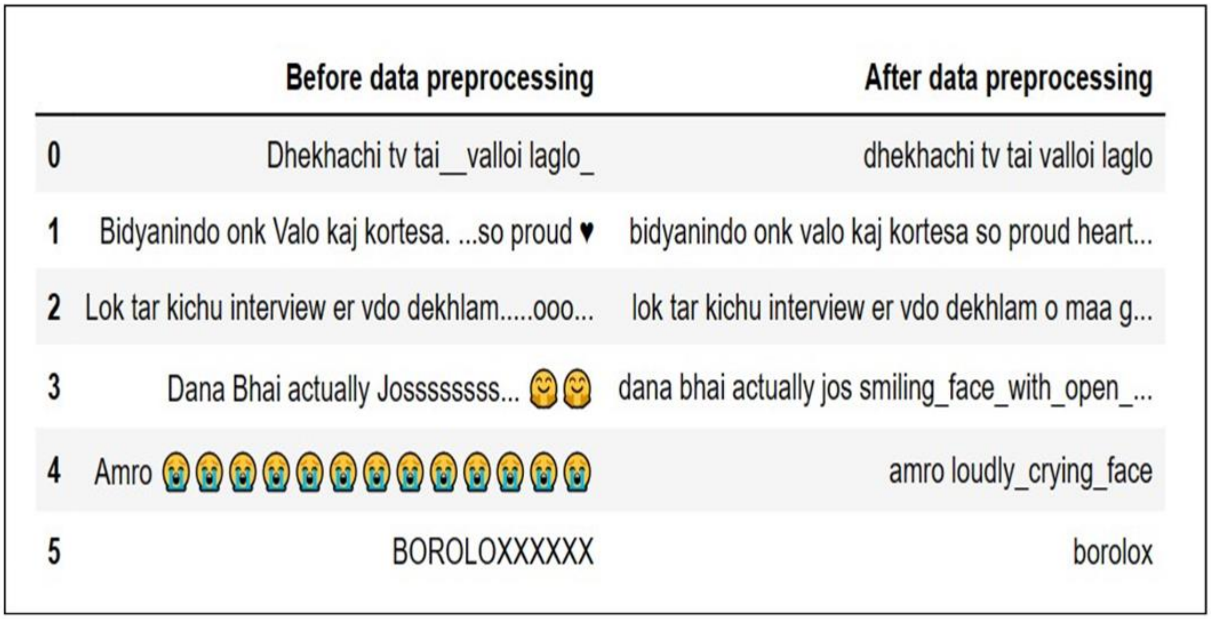
1. Handling missing data: The handling of missing data operations is carried out on the entire dataset. As missing data can bias the result of the model, If a row with a missing value is detected, the entire row is dropped.
2. Remove non relevant information: hyperlink, URL, email, name mention, hashtag, number, new lines, and extra space have no significant rules for detecting sentiment. So, these are removed from the comments.
3. Convert in lowercase: To avoid any case sensitivity all letters are converted into lowercase.
4. Remove punctuations: Punctuation as like ('!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~') has been removed from comments.
5. Convert emoji to Unicode CLDR: Each emoji has a unicode CLDR short name, like (😢 | crying face). For better performance in the model, emojis are converted into Unicde CLDR shortname.
6. Remove duplicate values: This operation is done in two steps. In the first step, if any word contains sequentially repeated letters more than twice, then the repeated letters have been removed. Such as (onnnnkkk-onk). In the second step, if any word is sequentially repeated, then the repeated word has been removed. Such as (sokal sokal egula valo lage na - sokal egula valo lage na).

3.3 Tokenization

After completing the data preprocessing steps to convert the raw dataset into a clean dataset, each data is converted into a list of tokens.

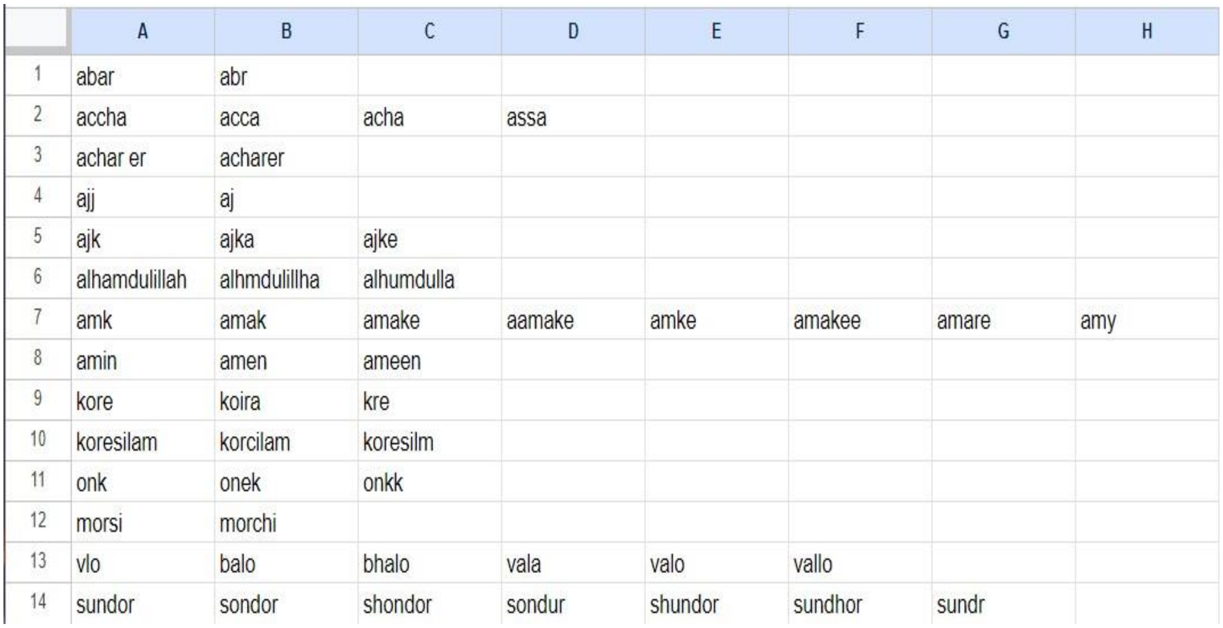
3.4 Normalization

In Bangla-English code mixed sentences there are various forms of words that are spelled differently but pronounced similarly which can be converted to one single form. As like (valo-vlo-bhalo, kharap-khrp etc). Another main things among words are, some people like to write short form of that word and some write the enlarge word. For the proposed model, normalization list have created manually with the words present in the dataset. The first word was used for replacing the variations of the spellings of the same Banglish word.



	Before data preprocessing	After data preprocessing
0	Dhekhachi tv tai__valloi laglo_	dhekhachi tv tai valloi laglo
1	Bidyanindo onk Valo kaj kortesa. ...so proud ♥	bidyanindo onk valo kaj kortesa so proud heart...
2	Lok tar kichu interview er vdo dekhlam.....ooo...	lok tar kichu interview er vdo dekhlam o maa g...
3	Dana Bhai actually Josssssss... 🤗🤗	dana bhai actually jos smiling_face_with_open_...
4	Amro 😭😭😭😭😭😭😭😭😭😭😭😭😭	amro loudly_crying_face
5	BOROLOXXXXXX	borolox

Fig. 3: Sample of Data Preprocessing



	A	B	C	D	E	F	G	H
1	abar	abr						
2	accha	acca	acha	assa				
3	achar er	acharer						
4	ajj	aj						
5	ajk	ajka	ajke					
6	alhamdulillah	alhmdulillha	alhumdulla					
7	amk	amak	amake	aamake	amke	amakee	amare	amy
8	amin	amen	ameen					
9	kore	koira	kre					
10	koresilam	korcilam	koresilm					
11	onk	onek	onkk					
12	morsi	morchi						
13	vlo	balo	bhalo	vala	valo	vallo		
14	sundor	sondor	shondor	sondur	shundor	sundhor	sundr	

Fig. 4: Sample of Created Normalization.

3.5 Feature Extraction

Each word are then transformed into numerical value using two different feature extraction method to evaluate which performs better for the model. Using TF-IDF and countVectorizer each unique token gets a feature index. Finally, each comment becomes a vector, and the weighted numbers of each vector represent the score of features.

3.6 Data Balancing

The dataset contains 887 Neutral, 559 Negative, and 609 Positive comments. This implies that the classes of the dataset are imbalanced which can effects the performance of the machine-learning model. To address the imbalanced data, there are several different approaches that can be taken, including both undersampling and oversampling techniques. Synthetic Minority Over-sampling Technique (SMOTE) has been used for this model to make a balance class dataset.

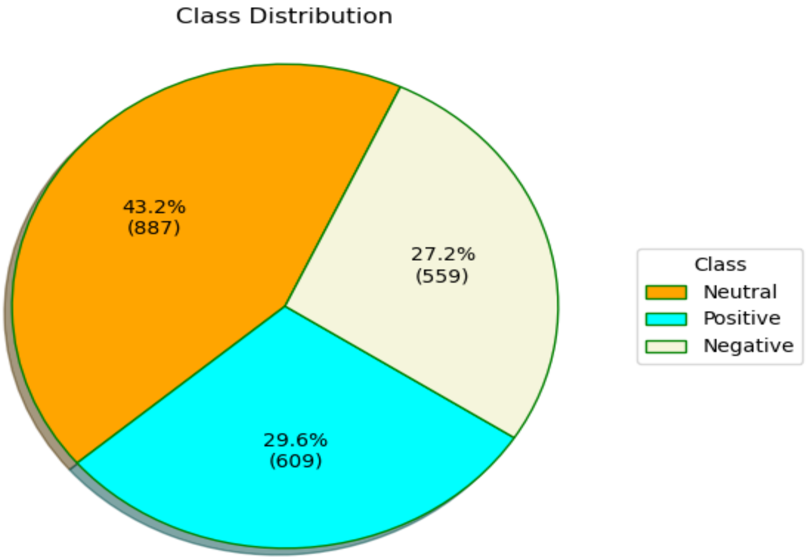


Fig. 5: Class distribution of the dataset.

3.7 Dataset Distribution and Classification

From the dataset 80% kept for train the model and 20% kept for evaluate the performance of model. For classifying comments on the basis of their sentiment into three classes: Positive, Negative and Neutral, nine machine learning algorithm has used; Bernoulli naive Bayes (BNB), Multinomial naive Bayes (MNB), Gaussian naive Bayes (GNB), Decision Tree Classifier (DT), K-Nearest Neighbour (KNN), Logistic regression (LR), Linear Support Vector Classifier (SVC), Support Vector Classifier (SVC) and random forests (RF).

4. RESULT ANALYSIS

The accuracy of several algorithm used to train the model dataset is shown in Figure 6. The barchart highlights that the algorithms obtained comparatively better accuracy using Term Frequency and Inverse Document Frequency (TF-IDF) rather than CountVectorizer feature extraction method. The highest accuracy, 85.7% obtained by Support Vector Classifier (SVC) using Term Frequency and Inverse Document Frequency (TF-IDF). The lowest accuracy obtained by K-Nearest Neighbour algorithm for both TF-IDF and CountVectorizer.

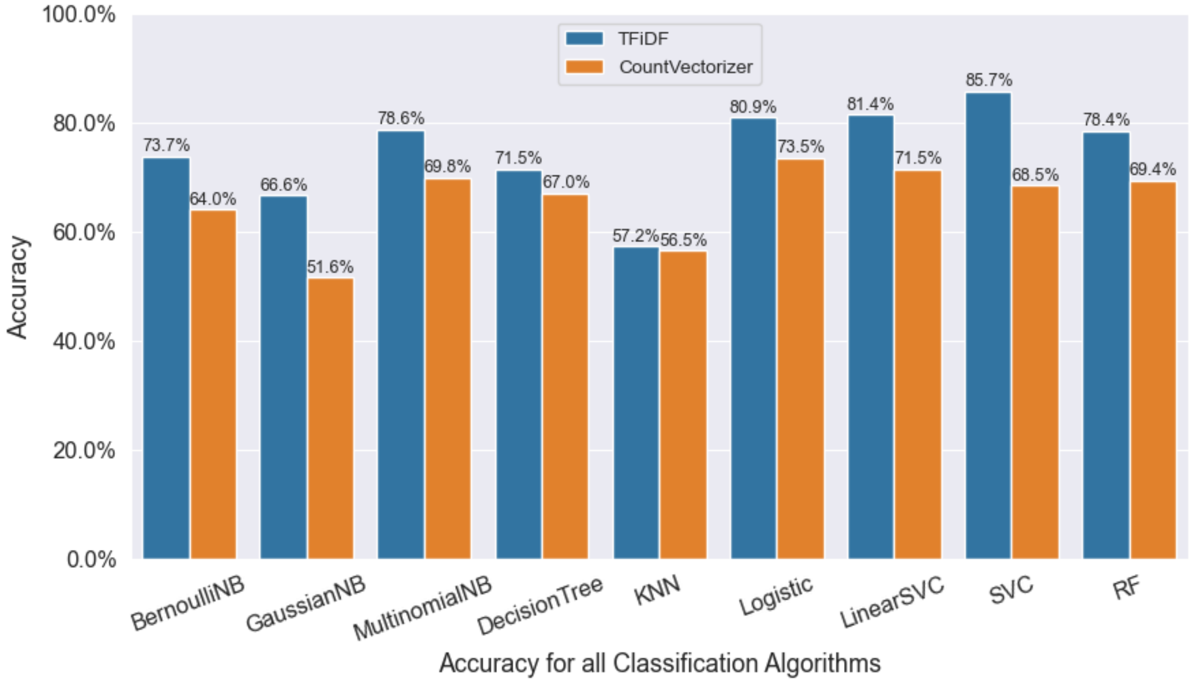


Fig. 6: Predictive accuracy barchart for TF-IDF and CountVectorizer.

For more depth analysis, three distinct metrics—Precision, Recall and F1-Score have been identified for implemented classification algorithms to evaluate and better comprehend the model's performance and predictive outcomes.

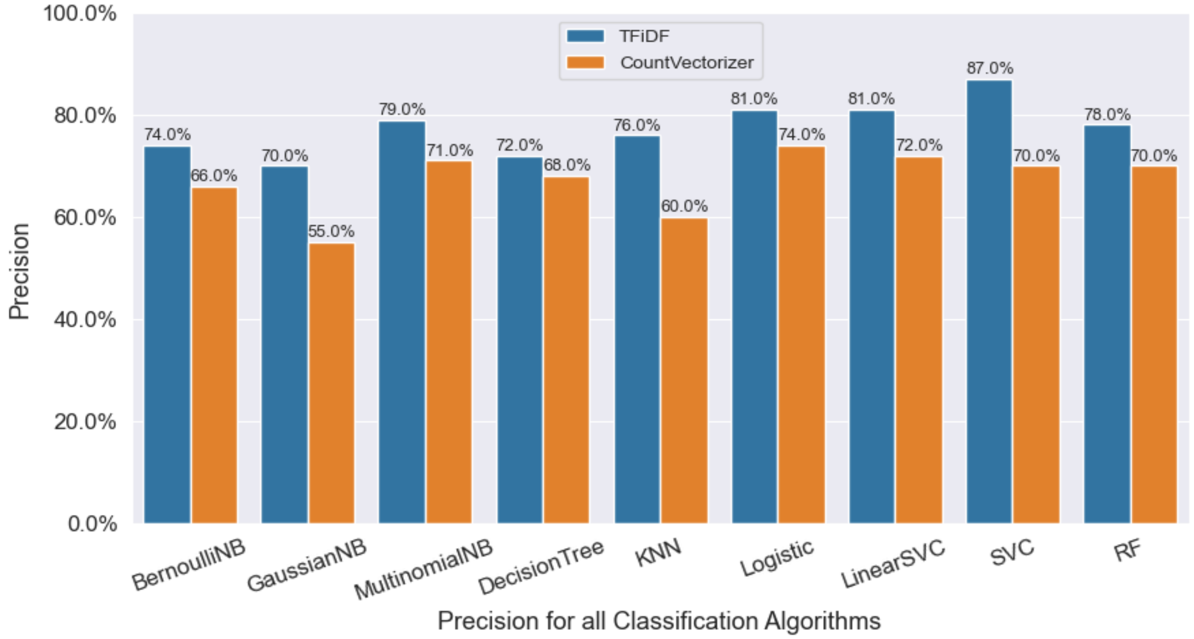


Fig. 7: Predictive Precision barchart for TF-IDF and CountVectorizer.

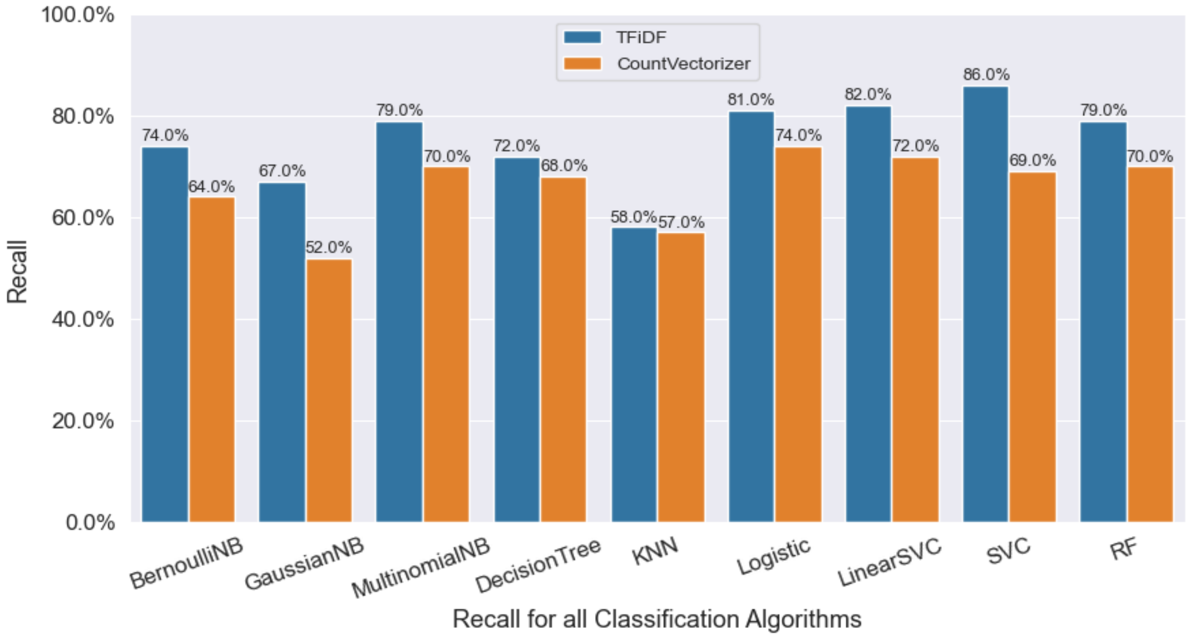


Fig. 8: Predictive recall barchart for TF-IDF and CountVectorizer.

Table 3: Precision, Recall and F1-Score of algorithms using Countvectorizer.

Classification	Precision	Recall	F1-Score
BernoulliNB	0.66	0.64	0.63
GaussianNB	0.55	0.52	0.49
MultinomialNB	0.71	0.70	0.69
DecisionTreeClassifier	0.68	0.68	0.66
K-Nearest Neighbors	0.60	0.57	0.52
LogisticRegression	0.74	0.74	0.73
Linear Support Vector Classifier	0.72	0.72	0.71
Support Vector Classifier	0.70	0.69	0.69
RandomForestClassifier	0.70	0.70	0.69

Table 3 and Table 4 shows the results for the nine considered classifiers using two different feature extraction method.

Table 4: Precision, Recall and F1-Score of algorithms using TF-IDF.

Classification	Precision	Recall	F1-Score
BernoulliNB	0.74	0.74	0.73
GaussianNB	0.70	0.67	0.64
MultinomialNB	0.79	0.79	0.78
DecisionTreeClassifier	0.72	0.72	0.71
K-Nearest Neighbors	0.76	0.58	0.52
LogisticRegression	0.81	0.81	0.81
Linear Support Vector Classifier	0.81	0.82	0.81
Support Vector Classifier	0.87	0.86	0.85
RandomForestClassifier	0.78	0.79	0.79

5. CONCLUSION & FUTURE WORK

This study presents a new technique for sentiment analysis of Bangla-English code-mixed language, including emojis, in social media communication. The study emphasizes the challenges of sentiment analysis in code-mixed language. It also highlights the importance of emojis for nuanced emotional interpretation, while acknowledging limited previous research. This study contributes to sentiment analysis by constructing a dataset, handling imbalanced data, creating a Bangla-English lemmatizer, and evaluating classification algorithms extensively. The results show that using the Term Frequency and Inverse Document Frequency (TF-IDF) feature extraction method with the Support Vector Classifier (SVC) achieved an accuracy of 85.7% and an F1 score of 85.0%.

This study proposes several potential directions for future exploration in sentiment analysis. These include expanding datasets, incorporating additional linguistic features, improving preprocessing techniques, and optimizing language variations to enhance effectiveness. Additionally, it aims to develop emotion recognition models that incorporate emojis, stickers, GIFs, and other elements to examine emotions in code-mixed text. These future research paths have the potential to not only improve the accuracy of sentiment analysis but also deepen our understanding of sentiment dynamics in code-mixing communities.

Acknowledgement

This project is financially supported by a research grant from Hamdard University Bangladesh.

References

1) Gaurav Singh, "Sentiment Analysis of Code-Mixed Social Media Text (Hinglish)", School of Computing, University of Leeds, Leeds, LS29JT, UK, doi.org/10.48550/arXiv.2102.12149.

2) Al-Azani S, El-Alfy ESM. Emoji-Based Sentiment Analysis of Arabic Microblogs Using Machine Learning. In: 2018 21st Saudi Computer Society National Computer Conference (NCC) [Internet]. Riyadh: IEEE; 2018 [cited 2023 Aug 18]. p. 1–6. Available from: https://ieeexplore.ieee.org/document/8592970/

3) GS. Solakidis et al. 2014. Multilingual sentiment analysis using emoticons and keywords. Proc. - 2014 IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol. - Work. WI-IAT 2014 2, (2014), 102–109.

4) AJ Shamal. 2019. Sentiment Analysis using Token2Vec and LSTMs: User Review Analyzing Module. (2019), 48–53.

5) Nusrath Tabassum and Muhammad Ibrahim Khan, "Design an Empirical Framework for Sentiment Analysis from Bangla Text using Machine Learning", 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019, doi.org/10.1109/ECACE.2019.8679347

6) L. K. Kaye, S. A. Malone, and H. J. Wall, "Emojis: Insights, affordances, and possibilities for psychological science," Trends in cognitive sciences, vol. 21, no. 2, pp. 66–68, 2017

7) "Emoji: The 21st Century Language", Divyansh Kandpal FEB 14, 2021, 18:57 IST, www.timesofindia.indiatimes.com/readersblog/dkwrites.

8) “World Emoji Day 2023: Celebrating the universal language of emojis and their digital impact”, Govind Choudhary, 17 Jul 2023, www.livemint.com/authors/govind-choudhary

9) “Emoji is the new universal language. And it’s making us better communicators”, Vyvyan Evans, Professor of Linguistics | Researcher, published, Aug 5, 2017, Linkedin, www.linkedin.com/pulse/emoji-new-universal-language-its-making-us-better-vyv-evans

10) “Emojipedia Shares Most Used Emojis of 2023 for World Emoji Day”, Andrew Hutchinson, Published July 17, 2023, www.socialmediatoday.com/news/emojipedia-shares-most-used-emojis-2023-world-emoji-day/688143/

11) P. K. Novak, J. Smailovic, B. Sluban, and I. Mozeti ´ c, “Sentiment of ˇ emojis,” PloS one, vol. 10, no. 12, p. e0144296, 2015.

12) https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2015

13) G. Abdulsattar et al. 2018. Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naive Bayes Classifiers.

14) Mohammed Alzubaidi and Farid Bourennani. 2021. Sentiment Analysis of Tweets Using Emojis and Texts. In 2021 The 4th International Conference on Information Science and Systems (ICISS 2021), March 17–19, 2021, Edinburgh, United Kingdom. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/ 3459955.3460608

15) Mahmud AA. Artificial Intelligence in Business Decision Making: A Study on Code-Mixed and Transliterated Bangla Customer Reviews. SSRN Electron J [Internet]. 2021 [cited 2023 Aug 25]; Available from: https://www.ssrn.com/abstract=3875534

16) Khan MdSS, Rafa SR, Abir AEH, Das AK. Sentiment Analysis on Bengali Facebook Comments To Predict Fan’s Emotions Towards a Celebrity. J Eng Adv. 2021 Jul 23;118–24.

17) Khan MdRH, Afroz US, Masum AKM, Abujar S, Hossain SA. Sentiment Analysis from Bengali Depression Dataset using Machine Learning. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) [Internet]. Kharagpur, India: IEEE; 2020 [cited 2023 Aug 18]. p. 1–5. Available from: https://ieeexplore.ieee.org/document/9225511/

18) Soumil Mandal, Sainik Kumar Mahata and Dipankar Das, “Preparing Bengali-English Code-Mixed Corpus for Sentiment Analysis of Indian Languages”, 11 Mar 2018, https://doi.org/10.48550/arXiv.1803.04000

19) M. Tareq, M. F. Islam, S. Deb, S. Rahman and A. A. Mahmud, "Data-Augmentation for Bangla-English Code-Mixed Sentiment Analysis: Enhancing Cross Linguistic Contextual Understanding," in IEEE Access, vol. 11, pp. 51657-51671, 2023, doi: 10.1109/ACCESS.2023.3277787.

20) Jamatia, S. D. Swamy, B. Gambäck, A. Das, S. Debbarma .(2020). Deep Learning Based Sentiment Analysis in a Code-Mixed English-Hindi and English-Bengali Social Media Corpus. International Journal on Artificial Intelligence Tools. doi:10.1142/s0218213020500141.

21) Patwa, P., Aguilar, G., Kar, S., Pandey, S., Pykl, S., Garrette, D., Gambck, B., Chakraborty, T., Solorio, T. and Das, A. 2020. SemEval-2020 Sentimix Task 9: Overview of SENTIment Analysis of Code-MIXed Tweets. Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)

22) A.M. MacEachren et al. 2010. Geo-Twitter Analytics: Applications in Crisis Management. Proc. 25th Int. Cartogr. Conf. (2010), 1–8.

23) B. G. Patra, D. Das, and A. Das. Sentiment analysis of code-mixed indian languages: An overview of sail code-mixed shared task @icon-2017. CoRR, abs/1803.06745, 2018.

24) Joshi, A. Prabhu, M. Shrivastava, and V. Varma. Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text. In Proceedings of the 26th International Conference on Computational Linguistics, pages 2482–2491, Osaka, Japan, Dec. 2016. ACL.

25) P. Nakov, Z. Kozareva, S. Rosenthal, V. Stoyanov, A. Ritter, and T. Wilson. SemEval-2013 Task 2: Sentiment analysis in Twitter. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the 7th International Workshop on Semantic Evaluation, pages 312–320, Atlanta, Georgia, June 2013. ACL.

26) J. Barnes, R. Klinger, and S. Schulte im Walde. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 2–12, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.