Data in Brief 48 (2023) 109067

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

BAAD: A multipurpose dataset for automatic Bangla offensive speech recognition

Md. Fahad Hossain[a,*], Md. Al Abid Supto[b], Zannat Chowdhury[b], Hana Sultan Chowdhury[b], Sheikh Abujar[b]

[a] Department of Computer Science and Engineering, Daffodil International University, Bangladesh
[b] Department of Computer Science and Engineering, Independent University, Bangladesh

ARTICLE INFO

Article history:
Received 29 September 2022
Revised 30 January 2023
Accepted 10 March 2023
Available online 24 March 2023

Dataset link: BAAD: A Multipurpose Dataset for Automatic Bangla Offensive Speech Recognition (Original data)

Keywords:
Offensive speech
Bangla offensive speech
Speech recognition
Multipurpose dataset

ABSTRACT

In spite of being the fifth most spoken native language in the world, Bangla has barely received any attention in the domain of audio and speech recognition. This article represents a speech dataset of Bengali Abusive Words with some non-abusive wors which are very close to the abusive ones. In this work, a multipurpose dataset is presented to recognize automatic slang speech for Bangla language, which was prepared by collection, annotation, and refinement of data. It consists of 114 slang words and 43 non-slang words with 6100 audio clips. For the collection of slang words, 60 native speakers and for non-abusive words, 23 native speakers participated who were, speaking in various dialects from over 20 districts of Bangladesh, and 10 university students participated to evaluate this dataset including annotation and refinements. Researchers can use this dataset to develop an automatic Bengali Slang speech recognition system, and also it can be used as a new benchmark for creating speech recognition-based machine learning models. This dataset can be enriched further, and some background noise in the dataset can be

* Corresponding author.
E-mail address: fahad15-9600@diu.edu.bd (Md.F. Hossain).
Social media: @fahad_hossain35 (Md.F. Hossain)

https://doi.org/10.1016/j.dib.2023.109067
2352-3409/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

used to simulate a more real-world scenario if desired. Otherwise, these noises could also be removed.

© 2023 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Specifications Table

Subject	Artificial Intelligence
Specific subject area	Signal Processing, Data related to nature, Bengali Abusive Language Recognition, Human Voice and Linguistics, Signal Processing.
Type of data	Audio
How data were acquired	We have a collection of 114 Bengali slang words, 80% of which were taken from a dataset on GitHub, and the remaining 20% of words are added manually which are commonly used slang and are not presented on GitHub. Along with this we also added 43 Non-abusive words that are close to being abusive. Data collection is done by survey using a google form requesting to provide recorded data with proper instruction like required data format, recording environment, etc, and by field level data collected from general public. When collecting data from both medium we have to make clear about our identity and reason behind collecting data. We also ensure informed consent for every speaker. In both cases, participants provided audio files and their information. Participants were instructed to use the following apps for recording voice data; "Easy Voice Recorder" for Android users, "Hokusai 2" for iOS users.
Data format	Raw WAV
Description of data collection	With the aim of collecting data in an organized and standard manner, an efficient data collection method was designed. Conditions for collecting data included that a participant should record natively in .wav format, where participants must take around 1 s pause in-between uttering each word while recording audio. When all the audio data was received, damaged and unusable recordings were filtered out which weren't working properly with our splitting algorithm. Bad data were weeded out to ensure chastity of the dataset, and good data were differentiated from bad data simply using two criteria; slang that was not split properly was ignored, for example, 2-3 slang words that got lumped together were dumped, and mispronounced slang words were also ignored. Dataset inclusion criteria includes, properly pronounced slang words by the user from the list of slangs, as well as, properly split whole words by our splitting algorithm.
Data source location	Institution: Independent University, Bangladesh (IUB) City/Town/Region: Dhaka Country: Bangladesh
Data accessibility	Repository name: Mendeley Data identification number: 10.17632/w24g8xn23c.3 Direct URL to data: https://data.mendeley.com/datasets/w24g8xn23c/3 Most of the slang words are collected from Github URL: https://github.com/rezacsedu/Bengali-Hate-Speech-Dataset

Value of the Data

- This dataset's data can be used to develop and evaluate machine learning and deep learning models for the detection of Bengali slang words. It can be used to build an automated system that can detect Bengali Abusive Language in the application of Bengali speech recognition system.
- Researchers can use this data to build an automated Bengali Abusive word detection system. Using this system, the public, especially minors, can be kept safe from exposure to abusive/harsh Bengali Slang terms.
- This data can be used as a foundation for building a bigger, more enriched dataset. It will also be useful for preliminary benchmarking for training/testing new ML models to detect Bengali abusive speech.

- This dataset is very much realistic as we considered some factors while collection. As the accent of different districts are different therefore the accent of the slang would also be different. We made an effort to provide information from all districts. Additionally, the accent on the same slang words appears to be different for men and women. So, we took data from both male and female.
- During data collection, the participants made sure that there is minimal background noise. But as background noise can be cleaned and removed whenever required, data with some background noise can be useful for simulating a more real-world scenario.

1. Data Description

Abusive words are words that use, contain, or are characterized by harshly or coarsely insulting language. Abusive words, at most times, come across as vulgar and offensive to other people, expressing negative, disrespectful connotations, a low opinion, or lack of respect. Usage of Abusive language often targets people of some ethnic or racial community, religious groups, or sexual and gender minorities.

Abundant research has been conducted on detecting Abusive Language in English, and resources and tools for detecting English Abusive words are widely available. But very little research work can be found focusing on low-resource languages like Bengali, despite being the fifth most spoken native language in the whole world. Today there are approximately 228 million native Bengali speakers and another 37 million second-language Bengali speakers, which makes Bengali the fifth most-spoken native language and the sixth most spoken language by the total number of speakers in the world. This is very unfortunate that with the prosperity of technology and social media, widespread usage of Abusive words and cyberbullying are being aided unintentionally, especially in developing nations like Bangladesh where cyber security laws are not well enforced. Cyberbullying and threats are on the rise, and oftentimes even end up with fatal outcomes. Cyberbullying and threat victims committing suicide are on the rise as well. Children are getting exposed to harsh Bengali slang terms from unregulated online content which is not suitable for their age. It affects their mental health. Hence, it is necessary to work on creating a safe and healthy 'Digital Bangladesh' for Bengali people, especially youngsters.

There is little text-based research work done focusing on detecting abusive words from low-resource languages like Bengali. These works are primarily focused on detecting online text comments with Bengali Abusive words. Among the very few works focusing on Bangla speech recognition include a benchmark dataset for news audio classification in Bangla, the first of its kind [1], and an audio-only emotional speech corpus for Bangla language, "SUBESCO" was developed, which can be used by the researchers for prosodic and emotional speech analysis, and in cognitive psychology experiments related to emotion expression [2]. There is neither any work found focusing on Voice Recognition paired with detecting Bengali Abusive words, nor is there any resource. This dataset could help minimize the victims and children getting exposed to abusive remarks on videos/audios, and content with Abusive Words in them respectively.

Furthermore, this dataset of carefully collected Bengali Abusive Words aims to work towards achieving this goal. The 'abusive words.xlsx' file contains a list of 114 slang words, which were used by the participants during their voice recording of the slang data. This file has 114 rows and 1 column, which is a list of slang words. The participants read through the list of slang words to record their voice data and submit it to us *via* google form/email.

80% of the slang was collected from GitHub [3] and 20% were manually added by authors which mostly contained district-specific common slang in Bangladesh. This list of slang words was verified by 10 university students.

The features of the slang text data in the 'abusive words.xlsx' file include slang from various districts of Bangladesh. The content of this list of slang contains common Bengali abusive words. Voice data includes 60 participants speaking in various dialects and voice audio-recorded data is recorded natively by participants in .wav format. Fig. 1 shows that 65% of the participants were male and 35% were female contributors. We could not completely discard a sizeable female data

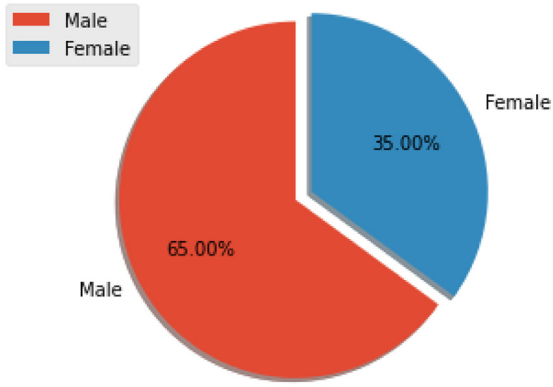

Fig. 1. Male to Female ratio in the audio dataset.

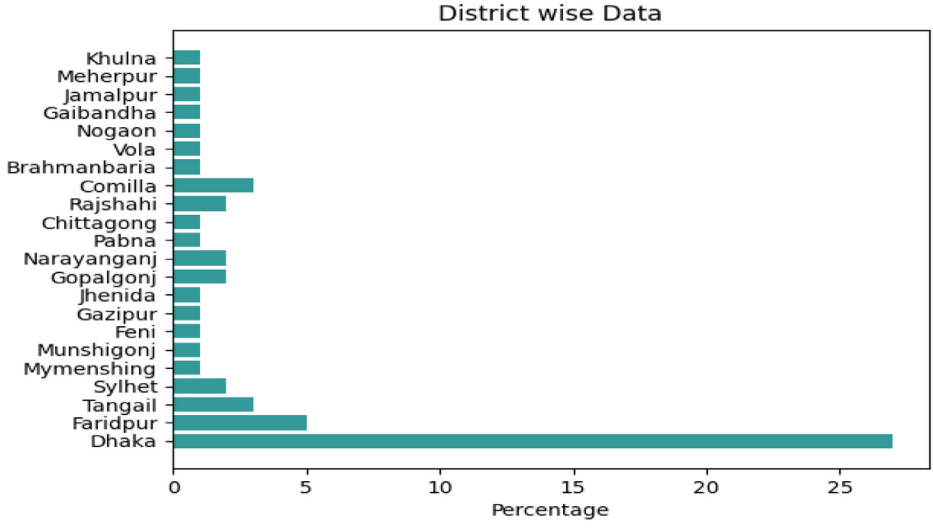

Fig. 2. District wise slang data distribution.

due to gender bias, as the female voice data is radically different from the male. Generally, the female vocal range sits on a higher pitch range and is more prominent with extreme emotions often connected with the usage of slang. This physical differentiation is important to emulate a realistic scenario when training machine learning models to detect slang. We believe that even after the observable gender bias in Fig. 1, there is still value that can be drawn from this scenario. It makes the dataset more realistic as the gender bias ratio shows a rough picture of the slang users in the population.

In different districts, people pronounce the same slang word differently as a result of which the accent of the people of different districts is not similar to hear. So, we tried to cover all of those pronunciations. From the collected personal information of the participants, Fig. 2 shows the district-wise slang data distribution, which indicates that the audio data consists of data in various dialects from over 20 districts.

Consequently, annotation and refinement of data, Fig. 3 shows the slang-wise total data collected using the 114 slang from 'abusive words.xlsx' and finally the amount of data in the dataset stands at 5277 slang audio clips.

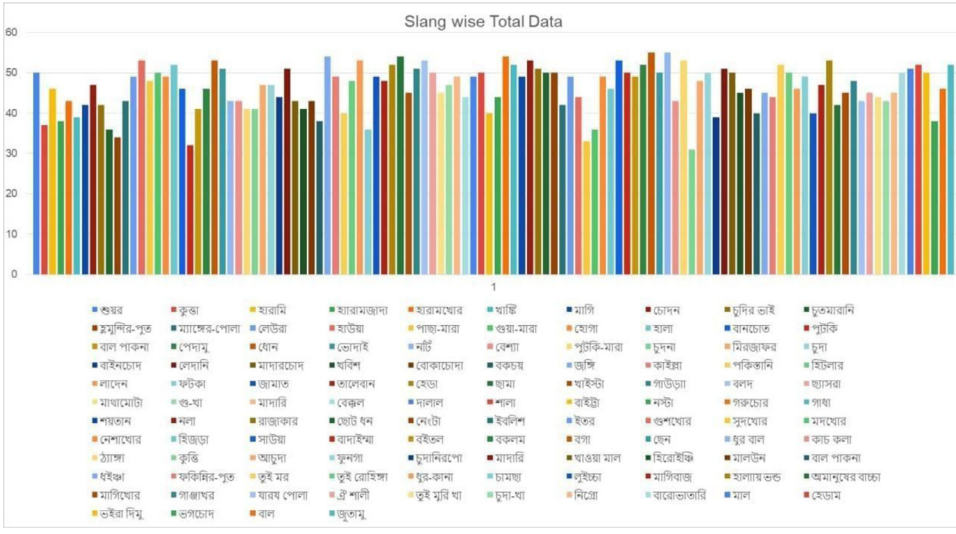

Fig. 3. Slang wise total data.

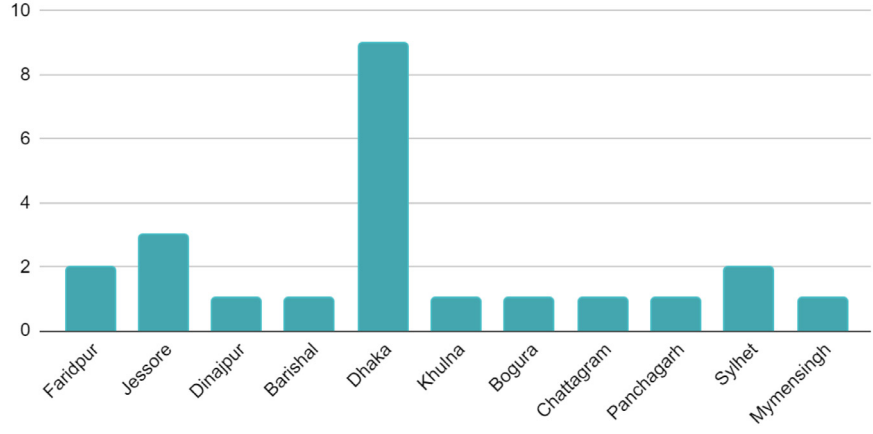

Fig. 4. District wise closest slang word data distribution.

In our dataset we have also included some data which are very close to those abusive words but are used commonly as not slang We have added “BAAD/ Abusive Colse Words.pdf “ file which includes the data which are not abusive but sounds very close to the abusive ones that are in the “abusive words.xlsx” file. These non slang voice data which are from 23 participants speaking in various dialects. And these voice data are recorded natively by participants in .wav format .Out of these 23 participant 16 are male and 7 are female contributors.We have added these non slang data so, the system does not detect these non-slang data as slang words. The distribution of district-specific slang data in Fig. 4 is based on participant’s personal data, and it reveals that the audio data includes data in diverse dialects from more than 20 districts.

After data annotation and refining we have shown the amount of each non-slang words in, Fig. 5 which displays the 43 closest to slang words from “Abusive/Slang Close Data.pdf “And lastly, there are a total of 823 non-abusive words in the collection of audio clips.

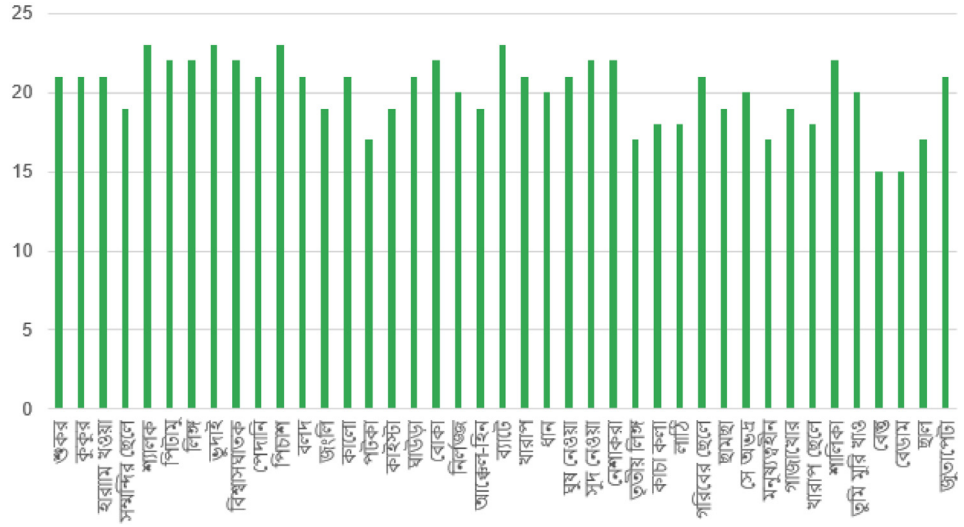

Fig. 5. Closest slang word wise total data.

Along with all the individual audio slang data from 60 speakers and non-slangdata from 23 speakers, we have also organized a metadata that we have refined personal information of each speaker, including gender, age, weight, height, and district.

The “BAAD_Unique_Speaker_information.xlsx” file and “BAAD_Close-Word_Speaker_info.xlsx ” file contains similar information which are all the unique speaker information. Where the first file contains information of the slang abusive words speaker and the second one is for the non-abusive words speakers. In both these excel sheetseach row contains unique speaker information with a “speaker_id” column. For the unique “speaker_id”, we encrypted each speaker’s name. “Speaker_id” column is followed by “gender”, “age”, “weight”, “height”, and “district” columns respectively, where each row holds this utilitarian personal information of that corresponding speaker.

“BAAD_Speaker_Information_WordWise.xlsx” and “BAAD Colse-Word Unique Speaker info.xlsx” contains word-wise metadata about each unique slang utterance audio file. In this excel sheets where each row is for a unique slang audio file; recognizable by its unique file path stored in the “path column”, it is accompanied by “word”, “speaker_id”, “gender”, “age”, “weight”, “height”, and “district” columns respectively.

2. Experimental Design, Materials and Methods

Handling huge amounts of data can be very troublesome unless the data is organized. For making our dataset, we have divided our workflow into two main parts named ‘data collection and ‘data Annotation’. Fig. 6 shows the workflow which was followed during data collection.

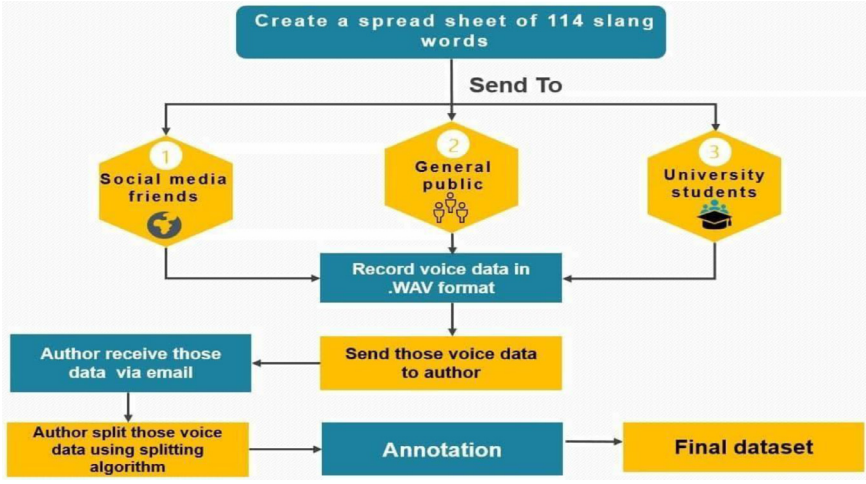

Fig. 6. Data collection workflow.

3. Data Collection

To operate the data collection phase smoothly and ensure the integrity and organized storage of data in our audio dataset, a google form and a data collection email was designed to collect audio data from the general public. The google form and email came with detailed instructions as to how a participant should record their audio data, including suggested reading pace, and ensuring low background noise and .wav format requirements to make sure data is of top quality. It also had instructions to use a specific app so that audio is natively recorded in .wav format. And the form also included an 'abusive words.xlsx' file, a list of slang words, which the participants refer to while reading out the slang words during recording. This google form and email were then sent to university students, the general public, and friends on social media. To retain all the participant's information as well, the google form and email were designed to ask each participant to input a few specific details about themselves like, Name, Age, Height, Weight, and Region. Participants input this data and then proceed to record their voice-recorded audio. Participants are instructed to use a specific app, namely 'Easy Voice Recorder' for Android, and 'Hokusai 2' for iOS devices, to record their voice recordings in .wav format natively on their mobile devices.

We have applied the same process for "Abusive/Slang Close Data.pdf" file for collecting the non abusive data which sounds close to the abusive ones Each participant is also instructed to rename their voice recorded .wav files to "Abusive.wav", for the abusive words and "non-abusive.wav" for non-abusive words then attach it to our google form and email.

4. Data Segmentation

Hossain. et al. worked on the segmentation of words from continuous Bangla voice by separating silent period and non-silent period with minimum dBFS value in audio clips [4]. Among their work, there was an algorithm developed for word segmentation. This algorithm was further modified to efficiently segment the slang audio data submitted by participants into chunks of 114 slang words by detecting pauses between uttering each word. Algorithm 1 shows the modified algorithm used to achieve this result.

Algorithm 1
Recursive splitting.

Input:	Input raw audio file, *A*
Output:	Create a copy of the audio, *B*
Step 1:	**FOR**
Step 2:	Set minimum silence length to, *n*
Step 3:	Set silence volume threshold to, *m*
Step 4:	Create a copy of the audio file, *c*
Step 5:	Using minimum silence length, split the audio file c into some instances
Step 6:	**FOR** every audio instance
Step 7:	**IF** instance volume is less than the silence volume, *m*
Step 8:	Store the starting time of silence
Step 9:	**ENDIF**
Step 10:	**ENDFOR**
Step 11:	Find and store all non- silence ranges using the list of starting time of silence
Step 12:	Split the audio, *B* into final instances on non-silence length and threshold is crossed
Step 13:	**IF** the number of instances is equal to the expected count
Step 14:	**BREAK**
Step 15:	**ENDFOR**

5. Data Annotation

Subsequently, after all the participant audio data were received, the 'Abusive.wav' and 'non-abusive.wav' audio files were inputted into the splitting algorithm respectively, which splits the files into chunks of 114 slang words and and 43 non-slang words is initially stored in the same participant folder.

With the aim of sorting these audio files into their corresponding slang folders, 114 separate slang folders were created and named with the slang word they are meant to store. 3 groups were made, and each group was tasked separately with the same goal; those groups were tasked to listen to each audio file, which was initially stored in chunks of 114 words in each participant's folder, and recognize the slang word uttered. If the utterance has any mistakes in pronunciation or is otherwise unusable, they would separate those files in a folder called "Bad Data". Contrastingly if the utterance is correct, they would put it in its right slang folder among all the 114 individual slang folders and 43 non-slang folders which were created earlier for them.

The supervisor performs a final check to make sure that all the audio files sit in their rightful corresponding slang folders. Ultimately all the chunk files are manually sorted into 114 individual slang folders and 43 non-slang folders, each folder is named after the slang or non-slang word it's going to contain.

Ethics Statements

Ethical approval was obtained from Research Ethics Committee of the Faculty of Science and Information Technology, Daffodil International University (REF-1002). Informed consent was obtained from all participants prior to participation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Data Availability

BAAD: A Multipurpose Dataset for Automatic Bangla Offensive Speech Recognition (Original data) (Mendeley Data).

Supplementary Materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2023.109067.

CRediT Author Statement

Md. Fahad Hossain: Conceptualization, Methodology, Software, Resources, Supervision, Data curation, Writing – review & editing; **Md. Al Abid Supto:** Conceptualization, Data curation, Investigation, Resources, Validation; **Zannat Chowdhury:** Writing – original draft, Data curation, Validation, Investigation; **Hana Sultan Chowdhury:** Visualization, Investigation, Resources, Validation; **Sheikh Abujar:** Resources, Writing – review & editing, Supervision.

References

[1] M. Rashid, M. Mahbub, M. Adnan, BAND: a benchmark dataset for bangla news audio classification, ACM Multimed. Asia 45 (2021) 1–6, doi:10.1145/3469877.3490575.
[2] S. Sultana, M. Shahidur Rahman, M. Reza Selim, M. Zafar Iqbal, SUST Bangla Emotional Speech Corpus (SUBESCO): an audio-only emotional speech corpus for Bangla, PLOS One 16 (4) (2021), doi:10.1371/journal.pone.0250173.
[3] Bengali Hate Speech Dataset, Github. 2020. https://github.com/rezacsedu/Bengali-Hate-Speech-Dataset. Accessed September 30, 2022.
[4] M. Fahad hossain, M Mehedi Hasan, H. Ali, S. Abujar, A continuous word segmentation of Bengali noisy speech, Adv. Intell. Syst. Comput. (2020) 525–534, doi:10.1007/978-981-15-7394-1_48.