Annihilate Hates (Task 4, HASOC 2023): Hate Speech Detection in Assamese, Bengali, and Bodo languages

Koyel Ghosh[1], Apurbalal Senapati[1] and Aditya Shankar Pal[2]

[1]Central Institute of Technology, Kokrajhar, Assam, India

[2]Indian Statistical Institute, Kolkata, India

Abstract

In today's world, social media can act as a tool for spreading hate towards a person or group based on their color, caste, sex, sexual orientation, political differences, etc. As social media continues to expand, the proliferation of hate speech is also surging at an alarming rate. Recently, Research on identifying hate speech in social media has gained significant prominence, with a specific need for investigations focused on languages other than English. The HASOC (Hate Speech and Offensive Content Identification) track intends to provide a platform for Hate Speech Detection since 2019 at FIRE (Forum for Information Retrieval Evaluation). HASOC 2023 is coordinating four tasks, with AH (Annihilate Hates, Task 4) being one of them. The AH task aims to develop and assess supervised machine learning systems on the three datasets. The three datasets presented for hate speech in three Indian languages (Assamese, Bengali, and Bodo) are collected from ™YouTube and ™Facebook comments. Each dataset is tagged with the binary classification (hate or non-hate) labels. In the Assamese language, 20 teams made 180 submissions, while 21 teams submitted 214 entries in the Bengali language, and for the Bodo language, 19 teams submitted a total of 175 submissions. The performance of the best classifiers for Assamese, Bengali, and Bodo are measured with the Macro F1 score of 0.73, 0.77, and 0.85, respectively. This article briefly summarizes the tasks, data development, and results. The variant of BERT architecture achieved the best performance in the task. However, other systems have also been successfully applied to the task.

Keywords

Hate Speech Detection, Binary Classification, Assamese, Bengali, Bodo, Machine Learning, Deep Learning, Transformers, BERT

1. Introduction

In addition to fostering friendships and facilitating information sharing, popular social media platforms such as ™Twitter, ™Facebook, and ™YouTube have also become platforms for cyberbullying and online harassment. These negative aspects can have severe consequences, including causing depression and inciting individuals to engage in violent actions, as evidenced in studies like [1, 2]. Instances of hate speech on these platforms have disrupted social and communal harmony on a global scale. Consequently, many countries have introduced increasingly complex regulations to address offensive online content, as discussed in [3] and [4]. This

Forum for Information Retrieval Evaluation, December 15-18, 2023, Goa, India

ghosh.koyel8@gmail.com (K. Ghosh); a.senapati@cit.ac.in (A. Senapati); adityashankarpal_r@isical.ac.in (A. S. Pal)

https://github.com/BrainLearns (K. Ghosh)

0000-0001-5347-4961 (K. Ghosh); 0000-0001-9124-2563 (A. Senapati); 0009-0001-0345-5428 (A. S. Pal)

situation has created a crucial need for automated methods to detect suspicious posts. It's worth noting that most research in this area has primarily focused on English and similar languages. On the other hand, Low-resource languages need more annotated datasets. Linguists have examined and characterized different manifestations of hate speech [5], while political scholars and legal authorities explore methods to govern online platforms and address problematic content while preserving the principles of free expression [6]. Algorithms are always getting better, and people are making lots of different sets of data for lots of other things, and they're studying them. Recently, Researchers made sets of data for many different languages [7] like English [8, 9, 10, 11], Greek [12], Portuguese [13], Danish [14], Mexican Spanish [15], and Turkish [16]. In Indian languages, hate speech dataset available is Hindi [17, 18, 19], Marathi [19], Bengali [20], Telegu [21], Tamil [21], Malayalam [21], and Kannada [21]. Having all these different data sets helps us understand how similar or different they are and how trustworthy they are.

In the HASOC 2023, four tasks (Task 1 - Task 4) in the research area of Hate Speech detection are proposed. Task 1 [22] focuses on identifying hate speech, offensive language, and profanity in different languages using natural language processing techniques. Task 2 [23], known as the Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL), addresses the challenge of identifying hate speech and offensive content in code-mixed conversations on social media. Code-mixed text includes multiple languages within a single conversation. The task is divided into two subtasks. Task 3 [24] aims to detect the various hateful spans within a sentence already considered hateful. A hate span is a set of continuous tokens that, in tandem, communicate the explicit hatefulness in a sentence.

This paper will provide an overview of Task 4, i.e., Annihilate Hates (AH), which contributes task-specific (Hate speech detection) low-resourced datasets on three languages: Assamese, Bengali, and Bodo. This AH dataset is version 3 and well updated of HS (version 2) [25], and NEIHS (version 1) [26, 27] datasets.

2. Related Forum and Dataset

The main obstacle in hate speech detection is the requirement for language-specific datasets. Constructing labeled datasets for hate speech in Indian languages is a laborious and intricate endeavor. It needs extensive groundwork and preprocessing tasks such as data cleaning and ensuring agreement among annotators. This section provides a concise overview of Indian datasets in languages like Hindi, Marathi, Bengali, Telugu, Tamil, Malayalam, and Kannada.

The HASOC challenge, organized by FIRE (Forum for Information Retrieval Evaluation)[1], has played a significant role in providing hate speech datasets in Indian languages like Hindi, Marathi, etc. HASOC comprises four subtracks, and the dataset distribution is in a tab-separated format. In 2019, the HASOC-Hindi dataset introduced three tasks as described in [17]. Subtask A is the initial task involving binary classification. Subtask B focuses on identifying the profanity or abuse within hate comments, a multiclass classification task. Subtask C is centered on determining whether the hate speech is targeted at a specific individual or if it's more general and untargeted. In the HASOC 2020 edition, two hate speech detection tasks were presented, as

[1] http://fire.irsi.res.in/fire/2022/home (Access on 30.10.2023)

mentioned in [18]. Subtask A involves binary classification, and Subtask B addresses multiclass classification. These tasks are accompanied by another Hindi dataset, expanding the research scope in this area. In 2021, HASOC published a Hindi dataset [19] with sub-tasks A and B again. Total Sixty-five teams submitted a total of six thousand and fifty-two runs. In HASOC-Marathi [19], the Marathi hate speech dataset with binary classification task. Authors [28, 29] experimented on HASOC datasets and analyzed the transformer-based model's performance in detail. BD-SHS [20] is a Bengali hate speech dataset with three levels: hate speech identification (binary classification, i.e., hate and not hate), identification of the Target of hate speech (multi-label classification, i.e., individual, male, female, and group), and categorization of hate speech types (multi-label classification, i.e., slander, call to violence, gender, religion). The authors [21] created several Indian datasets, i.e., Hindi, Telegu, Tamil, Malayalam, and Kannada, later performed monolingual, unbalanced splits, zero-shot cross-lingual, Few-shot, Joint training, pretraining and cross-dataset experiments on those datasets.

3. Task Description

In HASOC 2023[2], Task 4 is Annihilate Hates (AH) with three languages proposed in the research area of hate speech detection. These tasks offered all three languages: Assamese, Bengali, and Bodo. Figure 1 shows the Screenshot of Annihilate Hates (AH) Website[3].

3.1. Sub-task A: Hate Speech Detection in Assamese, Bengali, and Bodo (Binary)

Task 4 aims to detect hate speech in Assamese, Bengali, and Bodo languages. Each dataset (for the three languages) consists of a list of comments with their corresponding class (hate or offensive (*HOF*) or not hate (*NOT*)). Data is primarily collected from ™Facebook and ™YouTube comments. It is a binary classification task in which participating systems are required to classify the comments into two classes: *HOF* and *NOT*. Figure 2 shows the sample-tagged datasets of the AH-Assamese, AH-Bengali, and AH-Bodo.

4. Dataset Description

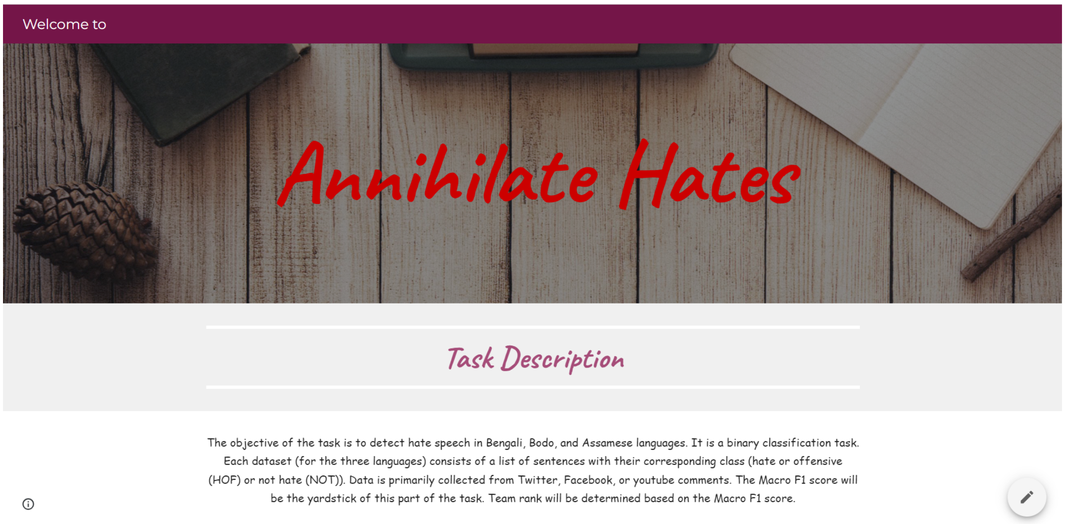
In this section, dataset collection, annotation, and analysis have been discussed for Task 4.

4.1. Dataset Collection

Our primary aim in constructing this dataset is to ensure its diversity, so we intentionally selected a few political, entertainment, and more ™Facebook pages and ™YouTube channels. We initiated the process by identifying contentious posts, often related to recent events, prominent figures such as politicians and actors, which had a higher likelihood of containing hate speech. Subsequently, we scrutinized the comments on these posts, seeking those primarily written in a

[2] https://hasocfire.github.io/hasoc/2023/index.html (Access on 30.10.2023)

[3] https://sites.google.com/view/hasoc-2023-annihilate-hates/home (Access on 30.10.2023)



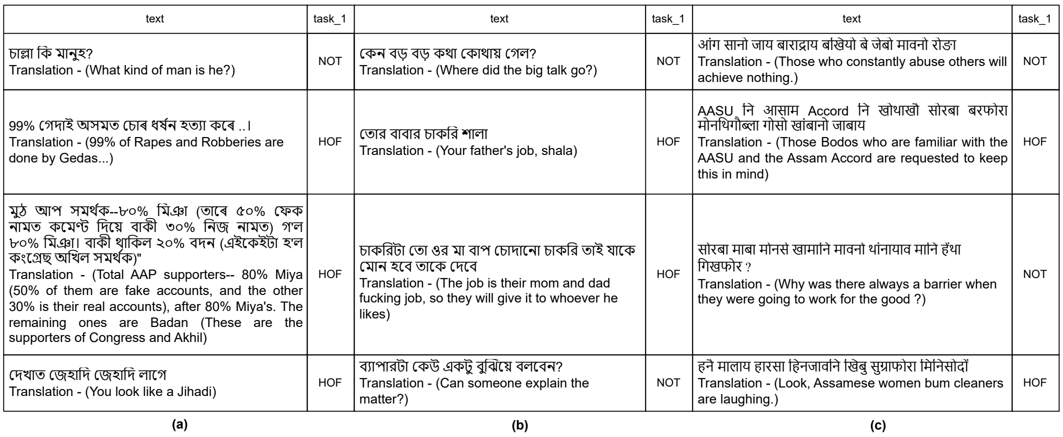
Welcome to

Annihilate Hates

Task Description

The objective of the task is to detect hate speech in Bengali, Bodo, and Assamese languages. It is a binary classification task. Each dataset (for the three languages) consists of a list of sentences with their corresponding class (hate or offensive (HOF) or not hate (NOT)). Data is primarily collected from Twitter, Facebook, or youtube comments. The Macro F1 score will be the yardstick of this part of the task. Team rank will be determined based on the Macro F1 score.

Figure 1: Screenshot of Annihilate Hates (AH) Website



text	task_1	text	task_1	text	task_1
চাল্লা কি মানুহ? Translation - (What kind of man is he?)	NOT	কেন বড় বড় কথা কোথায় গেল? Translation - (Where did the big talk go?)	NOT	आंग सानो जाय बाराद्राय बांखेयो बे जेबो मावनो रोङा Translation - (Those who constantly abuse others will achieve nothing.)	NOT
99% গেদাই অসমত চোৰ ধৰ্ষন হত্যা কৰে ..। Translation - (99% of Rapes and Robberies are done by Gedas...)	HOF	তোর বাবার চাকরি শালা Translation - (Your father's job, shala)	HOF	AASU नि आसाम Accord नि खोथाखौ सोरबा बरफोरा मोनथिगौब्ला गोसो खांबानो जाबाय Translation - (Those Bodos who are familiar with the AASU and the Assam Accord are requested to keep this in mind)	HOF
মুঠ আপ সমৰ্থক--৮০% মিঞা (তাৰে ৫০% ফেক নামত কমেণ্ট দিয়ে বাকী ৩০% নিজ নামত) গ'ল ৮০% মিঞা। বাকী থাকিল ২০% বদন (এইকেইটা হ'ল কংগ্ৰেছ অখিল সমৰ্থক)" Translation - (Total AAP supporters-- 80% Miya (50% of them are fake accounts, and the other 30% is their real accounts), after 80% Miya's. The remaining ones are Badan (These are the supporters of Congress and Akhil)	HOF	চাকরিটা তো ওর মা বাপ চোদানো চাকরি তাই যাকে মোন হবে তাকে দেবে Translation - (The job is their mom and dad fucking job, so they will give it to whoever he likes)	HOF	सोरबा माबा मोनसे खामानि मावनो थांनायाव मानि हैंथा गिखफोर ? Translation - (Why was there always a barrier when they were going to work for the good ?)	NOT
দেখাত জেহাদি জেহাদি লাগে Translation - (You look like a Jihadi)	HOF	ব্যাপারটা কেউ একটু বুঝিয়ে বলবেন? Translation - (Can someone explain the matter?)	NOT	हनै मालाय हारसा हिनजावनि खिबु सुग्राफोरा मिनिसोदों Translation - (Look, Assamese women bum cleaners are laughing.)	HOF

(a) (b) (c)

Figure 2: Samples tagged dataset of (a) AH-Assamese, (b) AH-Bengali, and (c) AH-Bodo

monolingual format, typically comprising 80-90%. We then conducted a manual assessment to determine whether these comments contained hate speech and categorized them accordingly. All the comments are collected using open source scrapper tools[4]. Ultimately, native speakers tagged the sentences as either *HOF* or *NOT*. Sentences that fell under the *HOF* category usually contained hate-related words and were considered hate-offensive statements. In contrast, sentences conveying formal information, suggestions, or questions were categorized as *NOT* sentences.

[4] https://github.com/kevinzg/facebook-scraper (Access on 30.10.2023)

4.2. Dataset Annotation

In dataset annotation, we share three separate CSV files with the three annotators. These files contain three columns: S. No. (serial number), text (comments), and task_1 (binary, i.e., *HOF* or *NOT*). The data for each language was tagged manually by three native speakers, young adults between 19 and 24. These annotators are students at the Central Institute of Technology in Kokrajhar, Assam, India. Their task involved manually classifying comments into two categories: those containing hateful (*HOF*) content and those that did not (*NOT*), using binary labels. The final decision was taken by consulting with a domain expert. Identifying hate speech is a subjective task, and it requires careful consideration. Consequently, we have established specific and rigorous guidelines to help define what qualifies as hate speech. These regulations are based on the community standards of ™Facebook[5] and ™YouTube[6]. The authors [21] follow the below-mentioned rules for the comments to be marked as hate, and we follow the scheme with updation. (a) *Profanity*: Comments that include profane language, curses, or vulgar words are categorized as hate speech. (b) *Sexual orientation*: Sexual attraction can be directed toward individuals of the opposite gender, the same gender, both genders or multiple genders. (c) *Personal*: Comments regarding one's fashion sense, choice of content, language selection, and related aspects. (d) *Gender chauvinism*: People are targeted in the comment because of their gender. (e) *Religious*: A person is criticized for their choice of religious beliefs and practices. For example, comments challenging the use of a turban or a burkha (the veil), (f) *Political*: Harassed a person based on political beliefs. For instance, bullying people for supporting a political party. (g) *Violent intention*: Containing a threat or call to violence in the comments.

Different annotators annotate the AH datasets, and the majority vote was considered; the annotation agreement calculated using κ (Kappa) coefficient is shown in Table 1. The problems and the level of disagreement need to be explored in the future.

Datasets	**κ statistics**
AH-Assamese	0.67
AH-Bengali	0.54
AH-Bodo	0.81

Table 1
κ statistics for all three datasets (Sub-task A)

4.3. Dataset Analysis

We summarize the key statistics of the AH dataset in Table 2. For the Assamese dataset, 2,955 comments are *HOF* out of 5,045. 641 comments are *HOF* out of 1,601 in the Bengali dataset which leads *NOT* is high. Out of 2,099, a total of 1,225 are *HOF* in the Bodo dataset. As a result, our Assamese and Bodo dataset is slightly skewed in favour of containing hate speech. Figure 3 shows the details of class distribution. In the training dataset, 4,036, 1,281, and 1,679 comments are present in the Assamese, Bengali, and Bodo datasets, respectively.

[5] https://web.facebook.com/communitystandards/ (Access on 30.10.2023)

[6] https://www.youtube.com/howyoutubeworks/policies/community-guidelines/ (Access on 30.10.2023)

Dataset	**HOF**		**NOT**		**Total**
	Train	**Test**	**Train**	**Test**	
AH-Assamese	2,347	608	1,689	401	5,045
AH-Bengali	515	126	766	194	1,601
AH-Bodo	998	227	681	193	2,099

Table 2
Class-wise distribution for AH-Assamese, AH-Bengali, and AH-Bodo datasets.

5. Result

The *macro* approach computes the F1 score individually for each class without considering the use of weights for aggregation. As a result, it imposes a more significant penalty when a system's performance is poor for minority classes. The selection of a specific F1 variant depends on the task's objectives and the label distribution in the dataset. Hate speech-related classification tasks often face class imbalance, making the macro F1 measure the suitable choice for evaluation.

For the system rum submission and evaluation of participants' experiments, we depend on the Kaggle platform. Figure 4 shows the Screenshot of the Annihilate Hates (AH) Kaggle Website for run submission. We provide separate Kaggle platforms Like Assamese[7], Bengali[8], Bodo[9] for participants to submit experimental runs.

Overall, 69 participants register for the task 4. In the Assamese task, 20 teams made 180 submissions, while 21 teams submitted 214 runs in the Bengali task, and for the Bodo task, 19 teams submitted a total of 175 runs.

The performance of the best classification algorithms for Assamese, Bengali, and Bodo are Macro F1 measures of 0.73, 0.77, and 0.85, respectively. The results for AH-Assamese, AH-Bengali, and AH-Bodo datasets are shown in Table 3, Table 4, and Table 5, respectively.

6. Methodology

This section discusses the systems utilized by the participants.

6.1. AH-Assamese

- *Chetona* [30] propose ensembling IndicBERT and Naive Bayes, along with synthetic data upsampling techniques (up-sample the training examples of each language by translating the examples from the other two languages to the given language.).
- *FiRC-NLP* [31] fine-tune the pre-trained XLM-RoBERTa-large model to get second position in the leaderboard.
- *TeamBD* [32] experimented with xlm-roberta-large (multilingual) along with ChatGPT3 augmentation.

[7] https://www.kaggle.com/competitions/annihilate-hates-assamese (Access on 30.10.2023)
[8] https://www.kaggle.com/competitions/annihilate-hates-bengali (Access on 30.10.2023)
[9] https://www.kaggle.com/competitions/annihilate-hates-bodo (Access on 30.10.2023)

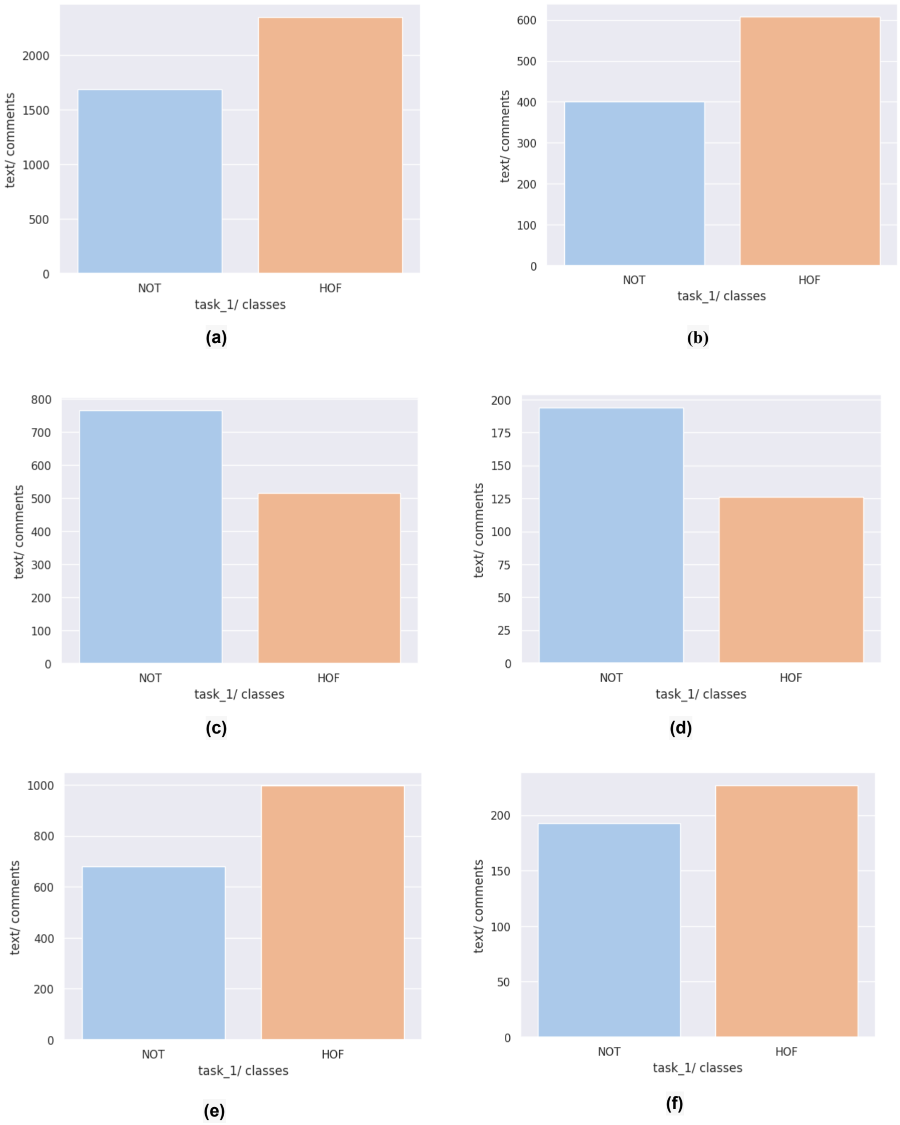
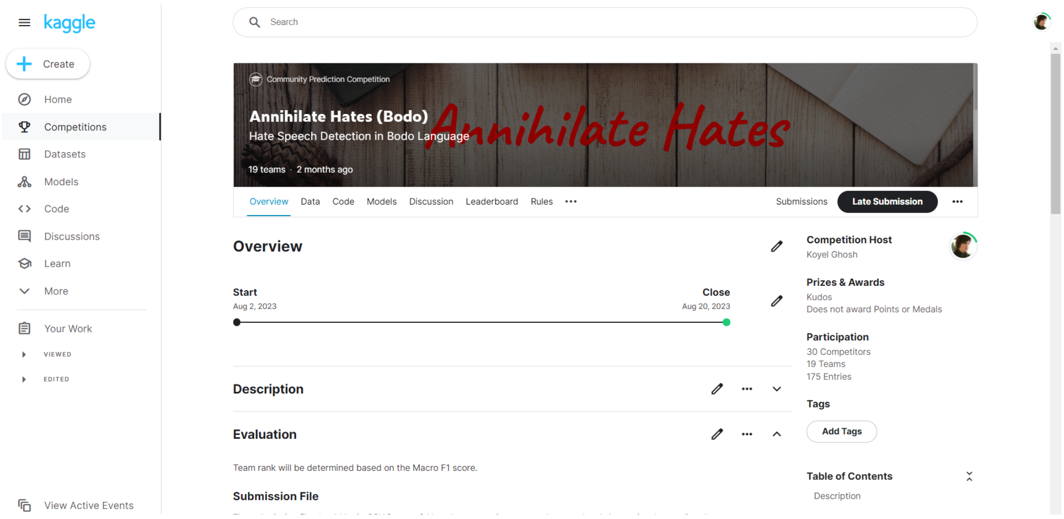


Figure 3: Class distribution of AH dataset with two classes (*HOF* and *NOT*) - a) AH-Assamese (train) and b) AH-Assamese (test), c) AH-Bengali (train), d) AH-Bengali (test), e) AH-Bodo (train) and f) AH-Bodo (test)

- *SATLab* [33] uses the LIBLinear L2-regularized logistic regression model (dual, -s 7) [42].
- *AI Alchemists* [34] fune-tuned XLM-RoBERTa model.
- *Sanvadita* [35] uses monolingual assamese-bert[10] and the multilingual indic-bert model[11].
- *Z-AGI Labs* [36] experiments with various multi-lingual transformer-based models for fine-tuning such as Bert-Base-Multilingual (Cased and Uncased), DistilBert-Base-

[10] https://huggingface.co/l3cube-pune/assamese-bert (Access on 30.10.2023)
[11] https://huggingface.co/ai4bharat/indic-bert (Access on 30.10.2023)



Community Prediction Competition

Annihilate Hates (Bodo)

Hate Speech Detection in Bodo Language

Annihilate Hates

19 teams · 2 months ago

Overview

Start
Aug 2, 2023

Close
Aug 20, 2023

Description

Evaluation

Team rank will be determined based on the Macro F1 score.

Submission File

Competition Host
Koyel Ghosh

Prizes & Awards
Kudos
Does not award Points or Medals

Participation
30 Competitors
19 Teams
175 Entries

Tags

Table of Contents
Description

Figure 4: Screenshot of the Annihilate Hates (AH) Kaggle Website for run submission.

Rank	Team name	Marco F1 score
1	Chetona [30]	0.7346
2	FiRC-NLP [31]	0.7251
3	TeamBD [32]	0.7222
4	SATLab [33]	0.7151
5	AI Alchemists [34]	0.7074
6	Sanvadita [35]	0.7064
7	Z-AGI Labs [36]	0.7052
8	Corgi	0.7044
9	JCT/ Avigail Stekel [37]	0.6988
10	Code Fellas [38]	0.6972
11	IRLab@IITBHU	0.6967
12	Komar99	0.6946
13	MUCS [39]	0.6883
14	Michal Stekel	0.6862
15	Chen876	0.6811
16	Ravens	0.6620
17	CNLP-NITS-PP [40]	0.5948
18	Team +1	0.4831
19	CIT TEAM	0.4684
20	InclusiveTechies	0.3468

Table 3
Result of task 4: Annihilate Hates (Assamese)

Rank	Team name	Marco F1 score
1	Sanvadita [35]	0.7702
2	FiRC-NLP [31]	0.7642
3	Z-AGI Labs [36]	0.7562
4	Daniil Orel	0.7507
5	TeamBD [32]	0.7349
6	AI Alchemists [34]	0.7257
7	Code Fellas [38]	0.7195
8	Chetona [30]	0.6785
9	SATLab [33]	0.6707
10	MUCS [39]	0.6683
11	JCT/ Avigail Stekel [37]	0.6649
12	Chen876	0.6603
13	Michal Stekel	0.6569
14	IRLab@IITBHU	0.6527
15	Komar99	0.6466
16	Ravens	0.6088
17	CNLP-NITS-PP [40]	0.6010
18	CHANDAN SENAPATI [41]	0.5062
19	Team +1	0.4709
20	CIT TEAM	0.3754
21	InclusiveTechies	0.3583

Table 4
Result of task 4: Annihilate Hates (Bengali)

Rank	Team name	Marco F1 score
1	SATLab [33]	0.8565
2	Komar99	0.8507
3	JCT/ Avigail Stekel [37]	0.8507
4	FiRC-NLP [31]	0.8484
5	Chetona [30]	0.8437
6	AI Alchemists [34]	0.8437
7	Ravens	0.8434
8	Chen876	0.8427
9	Michal Stekel	0.8378
10	MUCS [39]	0.8368
11	Code Fellas [38]	0.8351
12	Z-AGI Labs/ Nikhil Narayan [36]	0.8300
13	Corgi	0.8186
14	TeamBD [32]	0.7629
15	IRLab@IITBHU	0.7427
16	CNLP-NITS-PP [40]	0.6692
17	Team +1	0.4952
18	CIT TEAM	0.4152
19	InclusiveTechies	0.3148

Table 5
Result of task 4: Annihilate Hates (Bodo)

Multilingual-Cased, XLM-Roberta-Base, Muril-Base, and XLMIndic-Base (UniScript[12] and Multi-Script[13]) and got best result fine-tuning Bert Base Multilingual Cased model out of all experiments.

- *JCT/ Avigail Stekel* [37] develops different models using five classical supervised machine learning methods: multinomial Naive Bayes (MNB), support vector classifier, random forest, logistic regression (LR), and multi-layer perceptron. Their models were applied to word unigrams and/or character n-gram features. Their best model for the Assamese language is an MNB model with 5-gram features.
- *Code Fellas* [38] approaches which broadly involve Long Short Term Memory (LSTM) coupled with Convolutional Neural Networks (CNN) and pre-trained Bidirectional Encoder Representations from Transformers (BERT) based models like IndicBERT and MuRIL. Notably, their results showcase the effectiveness of these approaches, with IndicBERT achieving a remarkable F1 score for Assamese.
- *MUCS* [39], various experiments were carried out with different combinations of features (syllable n-grams, char n-grams, and fastText word embeddings) and different approaches (ML and FSL) to identify the given input. SVM trained with TF-IDF of syllable n-grams and TF-IDF of char n-grams, both in the range (1, 3), and this is their best model.
- *CNLP-NITS-PP* [40] experiments with CNN+FastText, CNN-BiLSTM+FastText/GLoVe, GPT-2, BERT, Logistic Regression. A CNN-based Binary Classification Model with FastText Embeddings outperforms their other systems.

6.2. AH-Bengali

- *Sanvadita* [35] uses pre-trained monolingual Bengali Sentence-BERT[14], Bengali-BERT models[15] and multilingual Indic Sentence-BERT [16].
- *FiRC-NLP* [31] utilizes XLM-RoBERTa-large model.
- *Z-AGI Labs* [36] utilizes pre-trained models for the experiments but achieves the highest score fine-tuning the csebuetnlp/banglabert pre-trained model.
- *TeamBD* [32] experiments with xlm-roberta-large model (multiingual).
- *AI Alchemists* [34] fune-tuned XLM-RoBERTa model.
- *Code Fellas* [38] fine-tuned MuRIL for Bengali to get the best experiment results out of all experiments done by the team.
- *Chetona* [30] applies the same for Bengali language as Assamese languages.
- *SATLab* [33] utilizes the same system as Assamese.
- *MUCS* [39], SVM trained with TF-IDF of syllable n-grams and TF-IDF of char n-grams both in the range (1, 3).
- *JCT/ Avigail Stekel* [37], their best model for Bengali is an MNB model with 6-gram features out of all experiments they performed, mentioned in the Assamese section.

[12] https://huggingface.co/ibraheemmoosa/xlmindic-base-uniscript (Access on 30.10.2023)
[13] https://huggingface.co/ibraheemmoosa/xlmindic-base-multiscript (Access on 30.10.2023)
[14] https://huggingface.co/l3cube-pune/bengali-sentence-bert-nli (Access on 30.10.2023)
[15] https://huggingface.co/l3cube-pune/bengali-bert (Access on 30.10.2023)
[16] https://huggingface.co/l3cube-pune/indic-sentence-bert-nli (Access on 30.10.2023)

- *CNLP-NITS-PP* [40], a CNN-based Binary Classification Model with FastText Embeddings outperforms the other systems.
- *CHANDAN SENAPATI* [41] implements the deep learning model LSTM.

6.3. AH-Bodo

- *SATLab* [33] utilizes the same system applied to the Assamese dataset.
- *JCT/ Avigail Stekel* [37], their best submission for Bodo is a LR with all word unigrams in the training set.
- *FiRC-NLP* [31] utilizes the XLM-RoBERTa-large model, the best submission among all their experiments.
- *Chetona* [30] applies the same system for the Bodo language as mentioned in the Assamese section.
- *AI Alchemists* [34] fune-tuned XLM-RoBERTa model.
- *MUCS* [39] trains SVM with TF-IDF of syllable n-grams and TF-IDF of char n-grams both in the range (1, 3) obtained the best macro F1 scores.
- *Code Fellas* [38] uses a BiLSTM model enhanced with an additional Dense Layer attaining an impressive F1 score for Bodo.
- *Z-AGI Labs/ Nikhil Narayan* [36] fine-tuned a pre-trained Bert Base Multilingual Cased model for Bodo.
- *TeamBD* [32] applies xlm-roberta-large model (multiingual).
- *CNLP-NITS-PP* [40] gets best result for Bodo dataset with Logistic Regression.

7. Conclusion

The submissions in the AH task (Task 4, HASOC 2023) have shown transformer-based pre-trained models to be the state-of-the-art approach for Hate Speech detection in the Assamese and Bengali datasets. However, the L2-regularized logistic regression model gives the best result for the Bodo dataset. Other deep learning models, like LSTM, CNNs, etc., also perform well on the given datasets. Upon reviewing the outcomes, the most suitable approach for hate speech detection depends on factors such as the language of the dataset, the level of classification detail, and the distribution of class labels. Balancing an imbalanced training dataset could impact the classification system's effectiveness. In the long run, the AH task aims to provide more low-resourced data with binary and multi-label classification tasks.

8. Acknowledgement

We thank Mr. Debarshi Sonowal, Mr. Abhilash Basumatary, and Ms. Bidisha Gogoi for their help in collecting and tagging the Assamese hate dataset. Additionally, we extend our thanks to Mr. Maharaj Brahma and Mr. Mwnthai Narzary for their valuable contributions in collecting and labelling the Bodo hate dataset. We also thank the FIRE and HASOC organizers for their support in organizing the track. We thank all participants for their submissions and their valuable work.

References

[1] M. L. Williams, P. Burnap, A. Javed, H. Liu, S. Ozalp, Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime, The British Journal of Criminology 60 (2019) 93–117. URL: https://doi.org/10.1093/bjc/azz049. doi:10.1093/bjc/azz049. arXiv:https://academic.oup.com/bjc/article-pdf/60/1/93/31634412/azz049.pdf.

[2] Z. Laub, Hate speech on social media: Global comparisons, Council on foreign relations 7 (2019).

[3] A. Nicholas, C. Ezeibe, The state, hate speech regulation and sustainable democracy in africa: a study of nigeria and kenya, African Identities (2020). doi:10.1080/14725843.2020.1813548.

[4] T. Quintel, C. Ullrich, Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond, Edward Elgar Publishing, 2019. Available at SSRN: https://ssrn.com/abstract=3298719.

[5] S. Jaki, T. De Smedt, M. Gwóźdź, R. Panchal, A. Rossa, G. De Pauw, Online hatred of women in the incels.me forum: Linguistic analysis and automatic detection, Journal of Language Aggression and Conflict 7 (2019) 240–268. URL: https://www.jbe-platform.com/content/journals/10.1075/jlac.00026.jak. doi:https://doi.org/10.1075/jlac.00026.jak.

[6] G. L. Casey, Ending the incel rebellion: The tragic impacts of an online hate group, Loyola Journal of Public Interest Law 21 (2019) 71. URL: https://heinonline.org/HOL/P?h=hein.journals/loyjpubil21&i=79.

[7] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 1–47. doi:10.1007/s10579-020-09502-8.

[8] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Predicting the type and target of offensive posts in social media, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1415–1420. URL: https://aclanthology.org/N19-1144. doi:10.18653/v1/N19-1144.

[9] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, Proceedings of the International AAAI Conference on Web and Social Media 11 (2017) 512–515. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14955. doi:10.1609/icwsm.v11i1.14955.

[10] I. Kwok, Y. Wang, Locate the hate: Detecting tweets against blacks, Proceedings of the AAAI Conference on Artificial Intelligence 27 (2013) 1621–1622. URL: https://ojs.aaai.org/index.php/AAAI/article/view/8539. doi:10.1609/aaai.v27i1.8539.

[11] Kaggle, Toxic comment classification challenge: Identify and classify toxic online comments (2017). URL: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

[12] Z. Pitenis, M. Zampieri, T. Ranasinghe, Offensive language identification in greek, CoRR abs/2003.07459 (2020). URL: https://arxiv.org/abs/2003.07459. arXiv:2003.07459.

[13] P. Fortuna, J. Rocha da Silva, J. Soler-Company, L. Wanner, S. Nunes, A hierarchically-labeled Portuguese hate speech dataset, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 94–104. URL: https://aclanthology.org/W19-3510. doi:10.18653/v1/W19-3510.

[14] G. I. Sigurbergsson, L. Derczynski, Offensive language and hate speech detection for danish, CoRR abs/1908.04531 (2019). URL: http://arxiv.org/abs/1908.04531. arXiv:1908.04531.

[15] M. Aragon, M. A. Carmona, M. Montes, H. J. Escalante, L. Villaseñor-Pineda, D. Moctezuma, Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets, 2019.

[16] Ç. Çöltekin, A corpus of Turkish offensive language on social media, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6174–6184. URL: https://aclanthology.org/2020.lrec-1.758.

[17] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th Forum for Information Retrieval Evaluation, FIRE '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 14–17. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584.

[18] T. Mandl, S. Modha, A. Kumar M, B. R. Chakravarthi, Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, in: Forum for Information Retrieval Evaluation, FIRE 2020, Association for Computing Machinery, New York, NY, USA, 2020, p. 29–32. URL: https://doi.org/10.1145/3441501.3441517. doi:10.1145/3441501.3441517.

[19] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, M. Zampieri, Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech, in: Forum for Information Retrieval Evaluation, FIRE 2021, Association for Computing Machinery, New York, NY, USA, 2021, p. 1–3. URL: https://doi.org/10.1145/3503162.3503176. doi:10.1145/3503162.3503176.

[20] N. Romim, M. Ahmed, M. S. Islam, A. Sen Sharma, H. Talukder, M. R. Amin, BD-SHS: A benchmark dataset for learning to detect online bangla hate speech in different social contexts, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 5153–5162. URL: https://aclanthology.org/2022.lrec-1.552.

[21] V. Gupta, S. Roychowdhury, M. Das, S. Banerjee, P. Saha, B. Mathew, h. p. vanchinathan, A. Mukherjee, Multilingual abusive comment detection at scale for indic languages, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 26176–26191. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/a7c4163b33286261b24c72fd3d1707c9-Paper-Datasets_and_Benchmarks.pdf.

[22] S. Satapara, H. Madhu, T. Ranasinghe, A. E. Dmonte, M. Zampieri, P. Pandya, N. Shah, M. Sandip, P. Majumder, T. Mandl, Overview of the hasoc subtrack at fire 2023: Hate-speech identification in sinhala and gujarati, in: K. Ghosh, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, Goa,

India. December 15-18, 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023.

[23] S. Madhu, Hiren Satapara, P. Pandya, N. Shah, T. Mandl, S. Modha, Overview of the hasoc subtrack at fire 2023: Identification of conversational hate-speech, in: K. Ghosh, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, Goa, India. December 15-18, 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023.

[24] S. Masud, M. A. Khan, M. S. Akhtar, T. Chakraborty, Overview of the HASOC Subtrack at FIRE 2023: Identification of Tokens Contributing to Explicit Hate in English by Span Detection, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, Goa, India. December 15-18, 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023.

[25] K. Ghosh, A. Senapati, Hate speech detection: an analysis of mono and multilingual transformer models with cross-language evaluation on hindi, marathi, bangla, and bodo language, Natural Language Engineering Accepted on 26.10.2023 (2023).

[26] K. Ghosh, A. Senapati, M. Narzary, M. Brahma, Hate speech detection in low-resource bodo and assamese texts with ml-dl and bert models, Scalable Computing: Practice and Experience 24 (2023) 941–955.

[27] K. Ghosh, D. Sonowal, A. Basumatary, B. Gogoi, A. Senapati, Transformer-based hate speech detection in assamese, in: 2023 IEEE Guwahati Subsection Conference (GCON), 2023, pp. 1–5. doi:10.1109/GCON58516.2023.10183497.

[28] K. Ghosh, D. A. Senapati, Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation, in: Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, Association for Computational Linguistics, Manila, Philippines, 2022, pp. 853–865. URL: https://aclanthology.org/2022.paclic-1.94.

[29] K. Ghosh, A. Senapati, U. Garain, Baseline bert models for conversational hate speech detection in code-mixed tweets utilizing data augmentation and offensive language identification in marathi, in: Forum for Information Retrieval Evaluation (Working Notes)(FIRE). CEUR-WS. org, 2022.

[30] S. Saha, M. Sullivan, R. Srihari, Hate Speech Detection in Low Resource Indo-Aryan Languages, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[31] M. S. Jahan, F. Hassan, W. Aransa, A. Bouchekif, Multilingual Hate Speech Detection Using Ensemble of Transformer Models, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[32] K. M. Jhuma, M. Oussalah, A. Singhal, Cross-Linguistic Offensive Language Detection: BERT-Based Analysis of Bengali, Assamese, & Bodo Conversational Hateful Content from Social Media, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[33] Y. Bestgen, Using Only Character Ngrams for Hate Speech and Offensive Content Identification in Five Low-Ressource Languages, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[34] C. Muhammad Awais, J. Raj, Breaking Barriers: Multilingual Toxicity Analysis for Hate Speech and Offensive Language in Low-Resource Indo-Aryan Languages, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[35] A. Joshi, R. Joshi, Harnessing Pre-Trained Sentence Transformers for Offensive Language Detection in Indian Languages, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[36] N. Narayan, M. Biswal, P. Goyal, A. Panigrahi, Hate Speech and Offensive Content Detection in Indo-Aryan Languages: A Battle of LSTM and Transformers, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[37] A. Stekel, A. Prives, Y. HaCohen-Kerner, Detecting Offensive Language in Bengali, Bodo, and Assamese using Word Unigrams, Char N-grams, Classical Machine Learning, and Deep Learning Methods, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[38] A. Reddy Gutha, N. Sai Adarsh, A. Alekar, D. Reddy, Multilingual Hate Speech and Offensive Language Detection of Low Resource Languages, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[39] P. M, R. K, A. Hegde, K. G, S. Coelho, H. L. Shashirekha, Taming Toxicity: Learning Models for Hate Speech and Offensive Language Detection in Social Media Text, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[40] G. Kalita, E. Halder, C. Taparia, A. Vetagiri, D. P. Pakray, Examining Hate Speech Detection Across Multiple Indo-Aryan Languages in Tasks 1 & 4, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[41] C. Senapati, U. Roy, Bengali Hate Speech Detection Using Deep Learning Technique, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[42] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, Liblinear: A library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.