Resources and Techniques for User and Author Profiling in Abusive Language (ResT-UP 2020), pages 2–8
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

An Indian Language Social Media Collection for Hate and Offensive Speech

Anita Saroj, Sukomal Pal
Department of Computer Science & Engineering
Indian Institute of Technology (BHU), Varanasi-221005, UP
anitas.rs.cse16@iitbhu.ac.in, spal.cse@iitbhu.ac.in

Abstract

In social media, people express themselves every day on issues that affect their lives. During the parliamentary elections, people's interaction with the candidates in social media posts reflects a lot of social trends in a charged atmosphere. People's likes and dislikes on leaders, political parties and their stands often become subject of hate and offensive posts. We collected social media posts in Hindi and English from Facebook and Twitter during the run-up to the parliamentary election 2019 of India (PEI data-2019). We created a dataset for sentiment analysis into three categories: hate speech, offensive and not hate, or not offensive. We report here the initial results of sentiment classification for the dataset using different classifiers.

Keywords: Twitter, Facebook, parliamentary Election, Hate Speech, Offensive

1. Introduction

Recent years have seen indiscriminate spread of offensive languages on social media platforms such as Facebook and Twitter. Hate speech and offensive posts day by day are growing on social media. People post messages or tweets, often targeting other people with hate and nasty words. Such messages often hurt people, causing at times immense psychological distress and mental trauma to users. Instead of bringing people together, it causes digital divide and social alienation to many. Such practices should be minimized, if can not be stopped entirely for reasons like maintaining the civility and decorum of any forum so that everyone can feel at home to participate. But often absence of any moderator to flag a post objectionable makes the job difficult. Efforts are, therefore, on to automatically detect the use of various forms of abusive languages in social networks, micro-blogs, and blogs so that prevention can also be thought of. Since manual filtering takes a lot of time, and since it can cause symptoms such as post-traumatic stress disorder to human annotators, several research efforts have made to automate this process (Zampieri et al., 2019a).

Few efforts have already been directed to create necessary datasets for automatic identification of offensive languages. The task is formulated as a supervised classification problem, where systems are trained for the presence of some form of abusive or offensive material. **Hate speech** in communication, is deemed to be harmful (individually or at a social level) based on defined 'protected attributes' such as race, disability, sexuality, etc., while **Offensive speech** is simply any communication that upsets someone.

Most of such datasets come from general domain and are in English. In this paper, we focus on in a particular domain with respect to space and time. During any election, when political rivalry reaches the summit, spread and use of obscene language also hit the ceiling. We consider the period of campaigning for general election of India 2019 and interactions of political candidates and people in social media. We present here the first domain-specific data of hate speech and offensive content identification on Parliamentary Election of India 2019 (PEI2019) data for two Languages, English and Hindi. The dataset is created from Twitter and Facebook posts during the Indian Election 2019. It comprises three tasks: a binary classification task, and two multi-class classifications.

Parliamentary Election of India (PEI data) data is especially inspired by two previous evaluation forums: HASOC FIRE 2019 (Mandl et al., 2019a) and SemEval 2019 (Zampieri et al., 2019a), and tries to leverage the synergies of these initiatives. There has been significant work in many languages, particularly for English, and the size of data is large. But there is no domain-specific data of hate speech and offensive content identification- which is the main motivation of making the PEI data. The size of PEI data is small but, we believe, enough to measure the performance of the classification models in Indian language hate speech dataset.

The primary purpose of the paper is to establish a lexical baseline for discriminating between hate speech and offensive speech on domain-specific data. Although some data for hate speech and offensive content identification are available,in English and other languages, there is no such dataset for the Indian language. Here we present a dataset of the Indian language, which is in Hindi and English dataset. We compare PEI 2019 data with two other datasets:

SemEval-2019 Task 6 and FIRE 2019 HASOC dataset.

The rest of the paper is organised as follows. In Sec 2., we do literature survey. Next, we describe the dataset in Sec 3.. We discuss the result in Sec 4.. Finally we conclude in Sec 6.

2. Related Work

Over the last few years, a few studies on hate speech and offensive content identification have been published. Different hate speech and offensive language identification problems are explored in the literature ranging from hate speech, offensive language, bullying content, and aggressive content. Below we discuss some of related works briefly.

2.1. Hate speech identification

Hate speech is a statement of intention to offend another and use harsh or offensive language based on actual or perceived membership to another group (Britannica, 2015). Malmasi and Zampieri (2017) adopted a linear support vector classifier with three groups of extracted features for these tests: word skip-grams, surface n-gram, and Brown cluster. They reported accuracy scores and established a lexical baseline for discriminating between profane and hate speech on the standard dataset (Malmasi and Zampieri, 2017).

2.2. Offensive language identification

While hate speech is targeted to a group of people based on their religion, caste, race, ethnicity or belief, offensive language such as insulting, harmful, derogatory, or obscene material is directed from one person to another and is open to others. Offensive language may be targeted or un-targeted. User-generated content on social media platforms such as Twitter often holds a high level of rough, harmful, or sometimes offensive language (Zampieri et al., 2019b). Increasing vulgarity in online conversations and user commentary have emerged as relevant issues in society as well as in science (Ramakrishnan et al., 2019). identified offensive tweets with an accuracy of 83.14 %, F_1-score 0.7565 on the real test data for the classification of offensive vs non-offensive.

The above tasks are related to that of cyber-bullying and aggressive contents and often differences are blurred. A post can contain one or many of the features above and can belong to many categories. However, we focused here on hate speech and offensive language identification tasks. The datasets mentioned were mostly in English and not domain-specific, but from general domain. As far as language specific collection is concerned, there has been probably the first task as HaSpeeDe 2018 [1] for Italian, PolEval 2019 and 2020 for Polish [2] and SemEval 2019 Task 5 that were domain-specific yet multi-lingual [3]. Here we build a domain-specific collection (political posts during election campaigns), and contain both English and Hindi posts. The vitriolic attacks become fierce as the campaign heats up and use of offensive languages nosedives to its nadir. We would like to see how the task of identifying hate and offensive language in such a collection and to gauge the extent of abusiveness in charged atmosphere.

3. Datasets

In India, the last parliamentary election was held from 11 April to 19 May 2019. During this event, we collected tweets and Facebook messages from social media in two languages Hindi and English. The data is used for training and testing in both hate speech and offensive language identification tasks. PEI data was annotated using a hierarchical three-level annotation model introduced in Zampieri et al. (2019) and Mandl et al. (2019).

3.1. Data Collection

We collected data from Facebook and Twitter during the parliamentary election 2019 of India. For Twitter, the data collection was done using the Twitter API with a tweepy Python library. The tweets collected from elected candidates' Twitter accounts and also collected with keywords #Twitter accounts name' and #Loksabha election, #election 2019, #loksabha election 2019 of India. For the hashtags, the tweets were between 11 April to 23 May 2019. For Facebook, we used the Facepager tool (Dr. Jakob Jünger, 2019) to capture messages. The collected tweets were in English, Hindi, and some other regional languages. For this study, we concentrated on tweets and messages in Hindi and English language. We collected more than ten thousand posts from Facebook and Twitter. Out of them, we found 20% tweets belonging to the hate speech and offensive content. Table 3.1. and Table 3.1. show some example of hate speech and offensive content in English and Hindi respectively.

3.2. Task Description

The dataset is created from Twitter and Facebook and distributed in a tab-separated format. The size of the data corpus is nearly 2000 posts for both English and Hindi separately. Figure 1 shows the categories of the post into different classes. The first stage categorization is Task A, and the second stage is Task B, and then, Task C as defined below.

- **Task A:** We focus on Hate speech and Offensive language identification for Hindi and English during the parliamentary election 2019 in India. Task A is a coarse-grained binary classification in which posts classify into two classes, namely: Hate

[1] http://www.di.unito.it/~tutreeb/haspeede-evalita18/index.html

[2] http://poleval.pl/

[3] https://www.aclweb.org/anthology/S19-2007/

Table 1: Tweets or Facebook messages from the PEI dataset, with their labels for each level of the annotation model of English.

Post	Label		
The Prime Minister talks about economic growth &progress. At the same time his colleagues talk about sending Bollywood stars to Pakistan!	NOT	-	-
NDTV features the Prime Minister's new improved BJP dream team for Karnataka. FRESH out of jail, MODI-FIED and REDDY to steal. #ReddyStingBJPExposed	HOF	HATE	UNT
West Bengal Chief Minister and Trinamool Congress supremo Mamata Banerjee on Monday called Prime Minister Narendra Modi the greatest danger for the country and said she will give her life to ensure that no riot takes place in the state.	HOF	OFFN	TIN

Table 2: Tweets or Facebook messages from the PEI dataset, with their labels for each level of the annotation model of Hindi.

Post	Label		
आज केरल और वायनाड के किसानों की समस्या लोक–सभा मे उठाया। उम्मीद है सर–कार इनका हल जल्द करेगी। Today the problem of farmers of Kerala and Wayanad was raised in the Lok Sabha. Hope the government solves these.	NOT	-	-
BJP और RSS के लोग धर्म की दलाली करते हैं।इनको न गाय से प्यार है, न धर्म से, इनको सिर्फ सत्ता से प्यार है–कानपुर देहात. People of BJP and RSS broke religion. They neither love cow nor religion, they only love power - Kanpur countryside.	HOF	HATE	TIN
बीजेपी की विचारधारा देश को बांटने की है, दलितों को कु–चलने की है, आदिवासियों को कुचलने की है, अल्पसंख्यकों को कुचलने की है, बीजेपी की उस विचारधारा के खिलाफ हम यहाँ खड़े ह The ideology of the BJP is to divide the country, crush the Dalits, crush the tribals, crush the minorities and are against that ideology of the BJP.	HOF	OFFN	TIN

and Offensive (HOF) and Non- Hate, or offensive (NOT).

- **Task B:** This is a fine-grained classification of Task A. Hate-speech and offensive posts from Task A further classified into three categories. **HATE** contains Hate speech content and **OFFN** contain offensive material and **NONE** not hate speech or not offensive.
- **Task C:** This one checks the type of offensive content. Only posts labeled as HOF in Task A are considered here. **Targeted Insult (TIN)** posts hold an abuse/threat to a person, group, or others. **Untargeted (UNT)** posts contain untargeted hate speech and offensive. Posts with general obscenity are considered not targeted, although they contain non-acceptable language.

3.3. Annotation

The annotation is done by three undergraduate students of Engineering whose first language is Hindi for speaking and writing, and they can speak and write English as well. The average score of inter-annotation agreement (Cohen's Kappa) for Task A is 0.87 for the English language and 0.89 for the Hindi language. Similarly, the average Cohen's Kappa for Task B and Task C are 0.85 and 0.89, respectively. We also evaluate Krippendorff's alpha which are 0.90, and 0.89 for English and Hindi respectively. Annotation labels for English and Hindi are shown in Table 1 and Table 2 and Figure 1 shows the hierarchy of annotations.

3.4. Data Summary

We consider Hindi and English language posts for hate speech and offensive content identification and some regional language. English and Hindi are the third and fourth most-spoken languages respectively, with Hindi having the largest number native-speakers in India [4]. Most of our collected posts in Hindi language, and some posts are code-mixed. The data can be used for multiple tasks in multi-way classification.

[4]https://en.wikipedia.org/wiki/Hindi

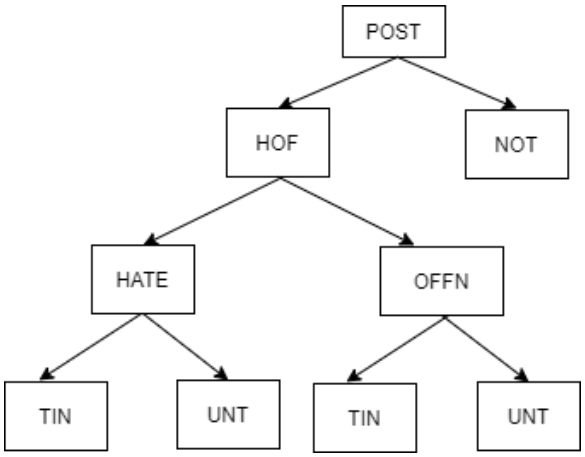

Figure 1: Process of the post or tweet annotation

Table 3: Distribution of labels combinations in PEI data.

Tasks	Labels			Total-Post	
Task A	HOF	NOT	-	Train	Test
Task B	HATE	OFFN	NONE	1519	488
Task C	UNT	TNT	NONE		

3.5. Data Preprocessing

Collected posts are first cleaned using the tweet preprocessing library[5] and several symbols like the Retweets (RT), Hashtags, URLs, Twitter Mentions, Emoji's and Smileys are removed. This pre-processed data also excludes the English stopwords (available in NLTK[6]) while tokenizing the sentences for the extraction of frequency-based feature extraction. Stopword removal and stemming are done on the terms. For prediction, the terms are represented by their tf-idf features considering each post as a document. These represented features are language independent and used for both Hindi and English. We did not use lemmatization, and any other lexical features that are language dependents.

3.6. Classifier

We use four machine learning classifiers: Multinomial Naive-Bayes (MNB), Stochastic Gradient Descent (SGD), Linear Support Vector Machine (Linear SVM), and Linear Regression (LR) for classification of Hate speech and Offensive content. The input for all the classifiers is in the form of tf-idf feature matrix, and output is a label for the categorical result. All the classifiers give different scores, as classifiers have different specialties.

3.7. Existing Data

For comparison, we also use similar data taken from other tasks. The first dataset of hate speech and offensive content is created by Davidson et al. (2017) and the second dataset is created by the HASOC track (FIRE 2019) (Mandl et al., 2019b). The SemEval-2019 Task 6 dataset is based on three subtasks, the Offensive Language Identification Dataset (OLID), which contains over 14,000 English tweets (Zampieri et al., 2019a). The HASOC track (FIRE 2019) is intended to encourage development in Hate speech identification for Hindi, German, and English language data. For English, HASOC 2019 has 5852 training instances, and 1153 instances for testing and for the Hindi language, the training corpus is 4665, and the testing corpus is 1318 (Mandl et al., 2019a).

[5]https://pypi.org/project/tweet-preprocessor/
[6]https://www.nltk.org/

4. Results

We begin by examining the accuracy of our tf-idf feature-based machine learning method. We first train the classifiers using tf-idf features. We perform classification on PEI 2019 data, SemEval 2018 task 6 (Zampieri et al., 2019a) and, FIRE 2019 task HASOC (Mandl et al., 2019b) for English datasets and compare our results with other standard benchmarks. We report classification performance of MNB, SGD, LR, and Linear SVM techniques in terms of precision (Pre), recall (Rec), F_1-score, and accuracy where their definitions considered are as given below.

1. Precision: It is the ratio of true-positives (TP) to the sum of true-positives and false-positives (FP).

$$Precision(P) = \frac{TP}{TP + FP} \tag{1}$$

2. Recall: It is the ratio of true-positives (TP) to the sum of true-positives and false-negatives (FN).

$$Recall(R) = \frac{TP}{TP + FN} \tag{2}$$

3. F_1-score: It is the balanced harmonic mean of precision and recall and used to have a composite idea of precision and recall.

$$F_1 = \frac{2 * R * P}{R + P} \tag{3}$$

4. $Macro\_F_1$: It is the average of per-class precision and recall scores over all classes. For each pair of classes, F_1 scores are computed and then arithmetic mean of these per-class F1-scores represent Macro-F_1.

5. $Weighted\_F_1$: It is the weighted version of the average F_1-scores where each class is weighted by the number of samples from that class.

6. Accuracy: It is the ratio of no. of correct predictions to the total number of original entities i.e.

$$Accuracy = \frac{\#\text{ correct predictions}}{\text{Total }\#\text{ test-instances}} \tag{4}$$

Table 4: Classifier performance on PEI-2019 for English data

Tasks	**Model**	MNB			SGD			LR			Linear SVM		
	Labels	Pre	Rec	F_1	Pre	Rec	F_1	Pre	Rec	F_1	Pre	Rec	F_1
Sub-task A	HOF	**0.97**	0.21	0.34	0.70	**0.43**	**0.53**	0.91	0.15	0.26	0.68	0.40	0.50
-	NOT	0.81	1.00	0.90	0.85	0.95	0.90	0.80	1.00	0.89	0.84	0.95	0.89
Sub-task B	HATE	**0.50**	0.03	0.05	0.32	0.10	0.15	1.00	0.03	0.05	0.35	0.10	0.16
-	NONE	0.78	1.00	0.88	0.84	0.96	0.89	0.79	1.00	0.88	0.83	0.97	0.89
-	OFFN	0.50	0.02	0.03	**0.85**	**0.61**	**0.71**	1.00	0.09	0.16	0.84	0.46	0.60
Sub-task C	NONE	0.80	0.99	0.88	0.84	0.93	0.88	0.79	0.98	0.88	0.84	0.93	0.88
-	TIN	**0.67**	0.15	0.24	0.55	0.39	**0.45**	0.64	0.13	0.22	0.55	**0.37**	0.44
-	UNT	0.00	0.00	0.00	0.80	0.29	0.42	0.00	0.00	0.00	0.75	0.21	0.33

Table 5: Classifier result of SemEval 2019 task 6 dataset at Precision, Recall, F-score and Accuracy.

Tasks	**Model**	MNB			SGD			LR			Linear SVM		
	Labels	Pre	Rec	F_1	Pre	Rec	F_1	Pre	Rec	F_1	Pre	Rec	F_1
Sub-task A	OFF	0.85	0.15	0.25	**0.92**	0.10	0.18	0.83	0.37	0.51	0.78	**0.46**	**0.58**
-	NOT	0.70	0.99	0.82	0.69	1.00	0.82	0.76	0.96	0.85	0.78	0.94	0.85
Sub-task B	GRP	0.00	0.00	0.00	0.00	0.00	0.00	**0.50**	0.03	0.06	0.48	0.05	0.10
-	IND	0.83	0.01	0.02	1.00	0.00	0.01	0.65	0.14	0.23	0.65	0.23	0.34
-	NULL	0.69	1.00	0.82	0.69	1.00	0.82	0.72	0.99	0.83	0.73	0.98	0.84
-	OTH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sub-task C	NULL	0.69	0.99	0.81	0.68	1.00	0.81	0.73	0.97	0.83	0.76	0.94	0.84
-	TIN	**0.77**	0.10	0.17	0.73	0.04	0.08	0.72	0.28	0.40	0.67	**0.39**	0.49
-	UNT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 6: Classifier result of FIRE 2019 task HASOC dataset at Precision, Recall, F-score and Accuracy.

Tasks	**Model**	MNB			SGD			LR			Linear SVM		
	Labels	Pre	Rec	F_1	Pre	Rec	F_1	Precision	Recall	F_1	Pre	Rec	F_1
Sub-task A	HOF	0.70	0.18	0.29	0.78	0.07	0.12	0.67	0.28	0.40	0.64	0.36	0.46
-	NOT	0.64	0.95	0.76	0.62	0.99	0.76	0.66	0.91	0.77	0.68	0.87	0.76
Sub-task B	HATE	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.03	0.05	0.29	0.06	0.10
-	NONE	0.62	1.00	0.77	0.63	1.00	0.77	0.64	0.98	0.78	0.65	0.95	0.77
-	OFFN	0.00	0.00	0.00	0.00	0.00	0.00	0.60	0.03	0.06	0.57	0.08	0.14
-	PRFN	0.86	0.04	0.07	0.78	0.12	0.20	0.78	0.12	0.20	0.79	0.18	0.29
Sub-task C	NONE	0.65	0.96	0.77	0.64	1.00	0.78	0.67	0.92	0.78	0.68	0.87	0.76
-	TIN	0.65	0.14	0.23	0.86	0.06	0.12	0.64	0.26	0.37	0.57	0.34	0.43
-	UNT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 7: Classifier result on testing dataset of PEI data

Task/Model	Sub-task A			Sub-task B			Sub-task C		
Model	Mac_f1	W_f1	Accuracy	Mac_f1	W_f1	Accuracy	Mac_f1	W_f1	Accuracy
Multinomial_NB	0.62	0.77	0.82	0.32	0.69	0.78	0.37	0.73	0.79
SGD	0.71	0.81	0.83	0.59	0.78	0.82	0.59	0.79	0.80
LR	0.58	0.75	0.81	0.36	0.70	0.79	0.37	0.73	0.79
Linear SVM	0.70	0.81	0.82	0.55	0.77	0.81	0.55	0.78	0.80

Table 8: Classifier result on testing dataset of SemEval 2019 Task 6 dataset

Task/Model	Subtask A			Subtask B			Subtask C		
Model	Mac_f1	W_f1	Accuracy	Mac_f1	W_f1	Accuracy	Mac_f1	W_f1	Accuracy
Multinomial_NB	0.54	0.63	0.71	0.33	0.59	0.69	0.21	0.57	0.69
SGD	0.50	0.61	0.70	0.30	0.56	0.68	0.21	0.57	0.69
LR	0.68	0.74	0.77	0.41	0.67	0.73	0.28	0.62	0.71
Linear SVM	0.71	0.76	0.78	0.44	0.70	0.74	0.32	0.65	0.72

Table 4 shows the result of PEI-2019 dataset for English. The machine learning models performed way better for PEI data than for the SemEval data-set. The reason is domain-specificity. While PEI dataset

Table 9: Classifier result on testing dataset of FIRE 2019 HASOC task dataset

Task/Model	Sub-task A			Sub-task B			Sub-task C		
Model	Mac_f1	W_f1	Accuracy	Mac_f1	W_f1	Accuracy	Mac_f1	W_f1	Accuracy
Multinomial_NB	0.53	0.58	0.65	0.21	0.49	0.62	0.33	0.56	0.65
SGD	0.44	0.51	0.62	0.24	0.51	0.63	0.30	0.52	0.64
LR	0.58	0.62	0.66	0.29	0.53	0.64	0.38	0.61	0.67
Linear SVM	0.61	0.64	0.67	0.35	0.56	0.64	0.40	0.62	0.66

Table 10: Classifier result of PEI-2019 dataset at Precision, Recall, F-score and Accuracy for Hindi data

Tasks	**Model**	MNB			SGD			LR			Linear SVM		
	Labels	Pre	Rec	F_1	Pre	Rec	F_1	Precision	Recall	F_1	Pre	Rec	F_1
Sub-task A	HOF	**0.85**	0.38	0.52	0.73	**0.64**	**0.68**	0.78	0.39	0.52	0.75	0.61	0.67
-	NOT	0.72	0.96	0.83	0.80	0.87	0.83	0.72	0.94	0.82	0.79	0.88	0.83
Sub-task B	HATE	0.33	0.02	0.04	0.59	0.34	0.43	0.57	0.17	0.26	0.63	0.36	0.46
-	NONE	0.64	0.99	0.78	0.76	0.93	0.83	0.68	0.98	0.80	0.73	0.96	0.83
-	OFFN	0.00	0.00	0.00	0.41	0.28	0.33	0.00	0.00	0.00	0.71	0.20	0.31
Sub-task C	NONE	0.73	0.98	0.84	0.81	0.89	0.84	0.75	0.98	0.85	0.82	0.91	0.86
-	TIN	0.79	0.30	0.44	0.67	0.59	0.63	0.81	0.35	0.49	0.72	0.60	0.66
-	UNT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 11: Classifier result on testing dataset of PEI Hindi data

Task/Model	Sub-task A			Sub-task B			Sub-task C		
Model	Mac_f1	W_f1	Accuracy	Mac_f1	W_f1	Accuracy	Mac_f1	W_f1	Accuracy
Multinomial_NB	0.67	0.71	0.74	0.20	0.50	0.64	0.42	0.69	0.74
SGD	**0.76**	**0.78**	**0.78**	**0.40**	0.67	0.70	0.49	0.76	0.77
LR	0.67	0.71	0.735	0.27	0.57	0.67	0.44	0.71	0.76
Linear_SV	0.75	0.77	**0.78**	**0.40**	**0.68**	**0.72**	**0.51**	**0.77**	**0.79**

is specific to election domain, SemEval contains posts from diverse domains. This affects the learning accuracy of the models, and hence PEI-2019 dataset performs better.

Table 5 and 8 show results of SemEval 2019 Task 6 dataset for English. The highest accuracy scores are 0.78, 0.74 and 0.72 for Subtask A, Subtask B and subtask C respectively.

We participated in FIRE 2019 (Saroj et al., 2019), and obtained the accuracy of XGBoost (81%) better than that of SVM (73%) for Subtask A (similar to Task A). The accuracy for Sub-task B and Sub-task C are the same for the XGBoost (80%). Table 6 and 9 show the FIRE HASOC English dataset results with accuracy 0.67, 0.64, 67 Subtask A, Subtask B and Subtask C respectively, where Mac_f1 is macro_f1 and W_f1 is weighted_f1.
The results above show that classification performance of PEI 2019 dataset is much better than the other dataset that are compared with for any of the techniques. In linear regression (LR), the macro-averaged F_1-score is 0.68 for SemEval 2019 dataset and 0.58 for the PEI 2019 dataset and FIRE 2019 dataset listed in Table 4, 5, and 6 respectively. The results of these experiments listed in Table 7, 8, and 9. Among the techniques, accuracy of the SGD classifier is the best among the three tasks (Task A, B, and C).

Table 10 and 11 show classification results for Hindi. The highest accuracy for Task A is 0.78 on SGD by linear SVM. For Tasks B and C, the highest accuracy are 0.72 and 0.79 respectively, again, by linear SVM.

5. Discussion

We found the highest accuracy in SGD classifier for all three subtasks in English data. For Hindi Linear SVM gives the best accuracy for all classes. LR gives better score in SemEval 2019 dataset compared to PEI 2019 and HASOC dataset. Multinomial NB, SGD, and Linear SVM give better F_1 score and accuracy in PEI 2019 dataset in all three subtasks than other datasets.

6. Conclusion

In this paper, we introduced a dataset for hate speech and offensive content detection in Indian language and Indian context. We tested a number of text classification techniques to recognize hate speech and offensive posts to validate our dataset: Multinomial Naive-Bayes, Stochastic Gradient Descent, Logistic Regression, and Linear Support Vector. The best results are achieved by Stochastic Gradient Descent (SGD), achieving 83% accuracy in three subtasks. We believe that tackling hate and offensive content in social media is a serious challenge and our PEI dataset will be useful, specifically in Indian context as it the first such dataset in any Indian language. In the future, we'd like

to apply domain adaptation and joint training from the parliamentary election 2019 of India.

7. References

Britannica, E. (2015). Britannica academic. *Encyclopædia Britannica Inc.*

Dr. Jakob Jünger, T. K. (2019). Facepager. *An application for generic data retrieval through APIs.*

Malmasi, S. and Zampieri, M. (2017). Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427.*

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019a). 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*, pages 14–17.

Mandl, T., Modha, S., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019b). Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages). In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation.*

Ramakrishnan, M., Zadrozny, W., and Tabari, N. (2019). UVA wahoos at SemEval-2019 task 6: Hate speech identification using ensemble machine learning. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 806–811, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Saroj, A., Mundotiya, R. K., and Pal, S. (2019). Irlab@ iitbhu at hasoc 2019: Traditional machine learning for hate speech and offensive content identification.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019a). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983.*

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019b). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.