SVM for Hate Speech and Offensive Content Detection

Shyam Ratan[1], Sonal Sinha[1] and Siddharth Singh[2]

[1]Department of Linguistics, Dr. Bhimrao Ambedkar University, India
[2]Centre for Transdisciplinary Studies, Dr. Bhimrao Ambedkar University, India

Abstract

This paper presents the system description of S_cube, which was submitted at the FIRE Shared Task 2021 on Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC). Our team submitted a system for Subtask 1 in two languages - English and Hindi, which has two different segments Subtask 1A and 1B for both languages. We experimented with the classic machine learning using Support Vector Machine (SVM). We discuss the system and its results with main findings for hate speech and offensive content identification in this paper. Our model achieves an F1 Score of 0.7563 at English Subtask 1A while the performance is worse for Hindi Subtask 1B (0.7195 F1).

Keywords

English, Hindi, SVM, Hate Speech, Offensive Language

1. Introduction

Communication on the internet has become a lot faster than anything in the world through various social media platforms like Facebook, Twitter, Whatsapp, Viber, Telegram and many more. Now, the concern is to check what kind of information and speech is being spread by users so that the social media platforms do not work as hotbeds for hate speech and offensive content. Therefore, a robust automatic filter system is required to sweep away these malicious contents.Hate speech and offensive content ranges over several issues, such as politics, religion, colour, gender, caste, ethnicity, etc. which holds the potential to polarise the society [1]. The benefit of anonymity and fake accounts on social media are major contributing factors for ease in bullying and the spread of hate speech and offensive languages at light speed.

Prominent efforts have been put to develop systems to secure the platforms (distinctively [2], [3], [4], [5], [6], [7], [8]). In addition to it, many shared tasks are being regularly organised for awareness and to come up with productive outcomes as automatic detection around the context of hate speech, aggression and offensive content [9], [10], [11], [1], [12], [13], [14].

One of its kinds in this horizon is FIRE 2021 shared task on Hate Speech and Offensive Content Detection in Indo-European Languages (HASOC 2021). In this paper, as part of the shared task, we elaborate on automatic hate speech and offensive content identification using SVM based system and its development for both segments of sub-task 1 in two languages - Hindi and English.

Forum for Information Retrieval Evaluation, December 13-17, 2021, India
shyamratan2907@gmail.com (S. Ratan); sonalsinha2612@gmail.com (S. Sinha); sidd435@gmail.com (S. Singh)

Table 1
The HASOC Dataset

	Train Sub-task 1A			Train Sub-task 1B					Test Set
	TOTAL	**HOF**	**NOT**	**TOTAL**	**HATE**	**OFFN**	**PRFN**	**NONE**	**TOTAL**
EN	**3,843**	2,501	1,342	**3,843**	683	622	1,196	1,342	**1,281**
HI	**4,594**	1,433	3,161	**4,594**	566	654	213	3,161	**1,532**

The remaining portion of the paper is divided into four sections. Section 2 discusses the used corpus size and its types for training and testing. Section 3 gives a detailed sketch of the conducted experiments for this task. Moreover, section 4 delivers the developed system's results and their error analysis with classified types of errors. Eventually, section 5 wraps up with the concluding notes.

2. Dataset

In order to direct the experiments for the identification/classification of hate speech and offensive language, we used the annotated twitter corpus for Hindi and English languages which were shared in the FIRE Shared Task HASOC 2021 [15]. An enumeration of the shared task corpus is given in Table 1. The corpus is labelled at two levels and they were presented as two segments in Subtask 1 for Hindi and English [16] given below -

1. **Subtask 1A:** In sub-task 1A, the corpus is annotated as HOF and NOT. HOF stands for hate speech, offensive language, and profane words while NOT is non hate and non-offensive content. Hence it is a binary classification task.
2. **Subtask 1B:** In Subtask 1B, fine-grained classification is offered for the identification of hate speech and offensive language. If the content is marked HOF in the Subtask 1A then it is marked as Hate Speech (HATE), Offensive (OFFN), and Profanity (PRFN) in this stage. Hence it is a three class classification task.

3. Experiments with SVM

We mainly experimented with SVMs classifier for Subtask 1A and 1B of Hindi and English corpus. We used the scikit-learn implementation of SVM ([17], [18] as cited in [1]). Support Vector Machines (SVMs) [19] are one of the most efficient classic machine learning models used for different kinds of text classification tasks. We experimented with binary and three-class problems with our basic objective of exploring the efficiency and productivity of SVMs for the detection of hate speech and offensive content.

In the case of our system, we experimented with SVM for both segments of Subtask 1 with the consecutive sets of features (given below in list 1, 2, and 3) and different C-values (0.001, 0.01, 0.1, 1, 5, 10) for working out the best model. Our classifier's best performances in both languages are given in Table 2 with n-gram features. Selection of these combination of word

Table 2
Comparison of character and word n-gram features for best SVM classifier

	Sub-task 1A		Sub-task 1B	
	EN	HI	EN	HI
Character n-grams	4, 5	4	4	3, 4, 5
Word n-grams	1, 2, 3	1, 2, 3	3	1

n-grams and character n-grams is based on best performances of system for Subtask1A and Subtask 1B.

1. Character n-grams features (trigrams to five-grams).
2. Word n-grams features (unigrams, bigrams and trigrams).
3. A systematic combination of diverse character n-grams and word n-grams features.

From the above experiments, we get that given the particular dataset, for English Subtask 1A feature of character five-gram and word trigram with C-value 10 gives the best performance. For Hindi Subtask 1A feature of character four-gram and word bigram with C-value 5 gives the best performance. For English Subtask 1B feature of character four-gram and word trigram with C-value 5 gives the best performance. For Hindi Subtask 1B feature of character four-gram and word unigram with C-value 10 gives the best performance. In the overall judgment of both Subtasks, the combination of character n-grams and word n-grams performed well for Subtask 1A in Hindi and English than Subtask 1B. Though, the score-wise improvement in all sub-tasks was very low for different features. Word n-gram features are widely effective in the case of Subtask 1A, which is binary classification and on the other side of three-class classification these are not very helpful for Subtask 1B.

4. Results and Error Analysis

In the collective results, our system performed best on the test set in Subtask 1A for English in comparison to Hindi (also Subtask 1A) and Subtask 1B for both languages. The macro F1 scores of all segments of Subtask 1 are present in Table 3.

Table 3
Macro F1 Score of Subtask 1

Subtask	Hindi Marco F1 Score	English Marco F1 Score
Subtask 1A	**0.7195**	**0.7563**
Subtask 1B	0.4513	0.5739

Our SVM classifier was placed at the 34th position in English Subtask 1A (the macro F1 score is 8 points below that of the topmost team), while it is placed at the 25th place in English Subtask 1B (macro F1 score being almost 9 points below the best performance team), it is placed at the 29th position in Hindi Subtask 1A (with an overall difference of 7 points in the micro F1 score of

the best team) and finally, it is placed at the 15th position in Hindi Subtask 1B (with a difference of almost 11 points in macro F1 score in comparison to the topper team). The performance comparison of our classifier and best classifier in all segments of Subtask 1 are summed up in the Figure 1.

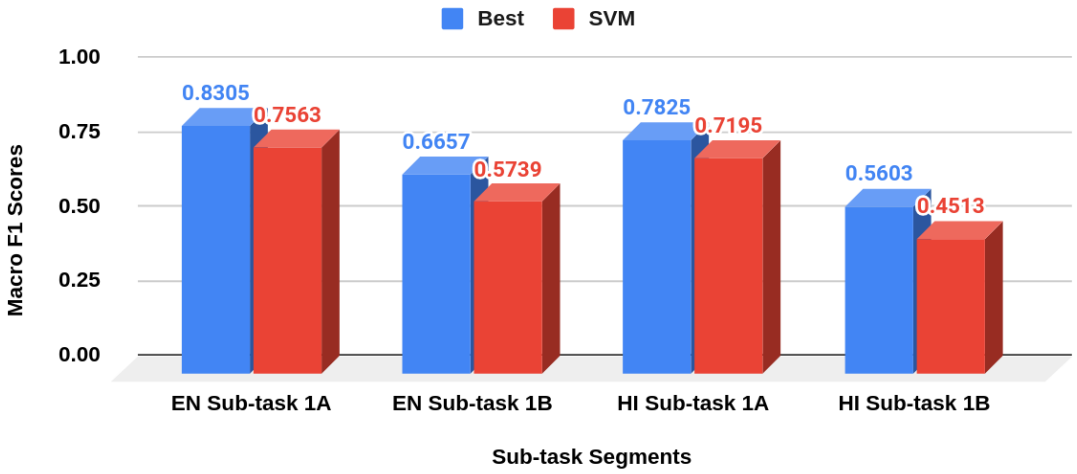

Figure 1: Performance of SVM classifier vis-a-vis the best classifier in Subtask 1

Apart from the results of our system for both languages, analysis of predicted errors on test data and its explanations are also most important. It is quite visible that the system we have developed has high precision and low recall in all Subtasks for Hindi and English.

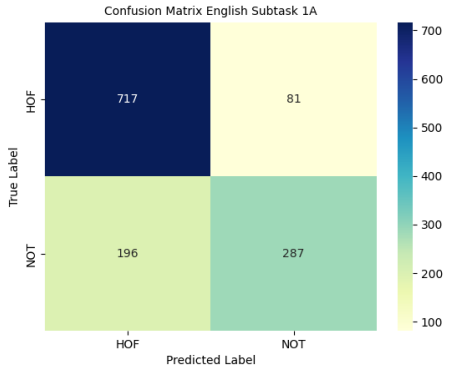

Figure 2: Confusion Matrix for English sub-task 1A

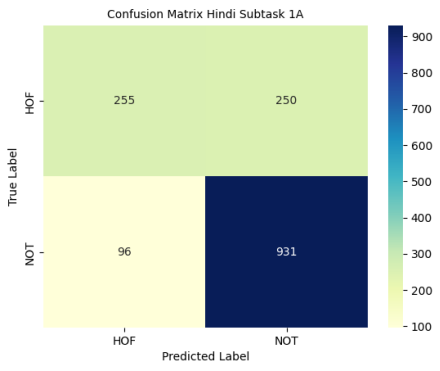

Figure 3: Confusion Matrix for Hindi sub-task 1A

In the comparison of predicted labels in sub-task 1A for both languages. Our system predicted the values of HOF class in English Subtask 1A (see Figure 2) are much higher in numbers than

the Hindi Subtask 1A (see Figure 3), which is quite opposite for NOT class in both segments of this Subtask. The performance of the system for different classes in different Subtasks is due to the sampling size of training sample data. Here, In Subtask 1A the proportion of both classes (HOF and NOT) were higher individually in English and Hindi.

Likewise, in three-class classification Subtask 1B the system performed well for some classes and predicted PRFN and NONE well in comparison of HATE and OFFN classes in English (see Figure 4). In this Subtask of Hindi (see Figure 5), NONE class is produced adequately good in numbers than the other classes (PRFN, OFFN, and HATE). The earlier trend of the proportion of training sample data is followed here in the case of three-class classification, where the 65% are PRFN and NONE classes of whole proportion in English, which is opposite in Hindi where PRFN is much lesser and OFFN, HATE are subsequent in numbers than NONE class. Another basis of the lower performance of the system in different Subtasks for both languages is the structure and morphological features of both languages, where the structure of Hindi is a little bit complex with a good number of morphological features.

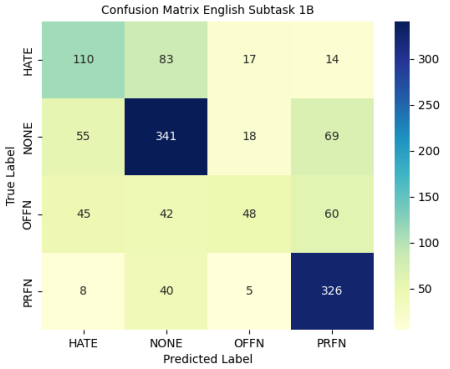

Figure 4: Confusion Matrix for English Subtask 1B

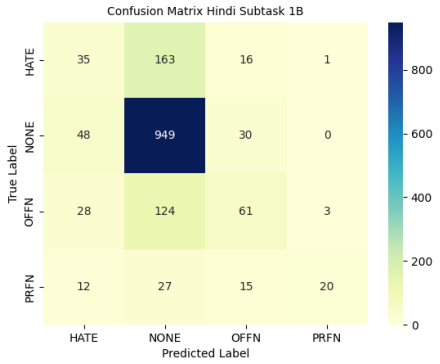

Figure 5: Confusion Matrix for Hindi Subtask 1B

In error analysis, some different types of errors are classified on the basis of gold labels and system predicted labels in both languages. These are satire/sarcasm, slogan, coined and aggressive lexical items, idiomatic expressions, quotes, and code-mixed data, etc. Broadly, these error types are predicted in the form of lexical features for all segments of Subtask 1 represented in Table 4.

5. Conclusion

The paper deals with a detailed description of the S_Cube system which is developed for HASOC at FIRE 2021. Results of the experiment show that SVM extrapolates a cut above for the binary classifier task in Subtask 1A, effective in cases of uneven corpus too, which is far opposite in the case of the three-class classifier. SVM is able to achieve low recall (but high precision) for all Subtasks in both languages. We also observed that, the lower performance in Subtask 1B could be broadly ascribed to the uneven corpus and the lack of ample training sample size for

Table 4
Error classification with types

Error Types	Hindi & English Examples	Translation & Explaination	Gold Label - Predicted Label
Satire / Sarcasm	1. vodaafon ne ek kuttaa paalaa thaa bhut fems huaa fir mukesh anbaani ko shauk kdha. 2. Kangana did a terrible mistake of pointing the mistakes of supreme leader !! Betrayal & amp	1. Vodafone had raised a dog, it became very famous then Mukesh Ambani was fond of it.	NOT - HOF, OFFN - HATE
Slogan	srkaar maun jntaa preshaan	The government is silent, the public is upset	NONE - HATE
Code-mix data	1. fattu hain bjp vaale. 2. In this may day we want a new fresh govt. not like this feku govt.	1. People of BJP (Bhartiya Janta Party)[1] are Coward. 2. We want new government in this may not like this government who makes false promises	HOF - NOT, HATE - NONE
Aggressive Lexical Items	aashutos tu vaakyi gadhaa hai	Ashutosh, you are a actual donkey	OFFN - NONE
Idiomatic Expressions	jaisi krni vaisi bhrni	As you sow, so you shall reap	NOT - HOF
Coined Lexical Items	godi midiyaa nmaajvaadi paarti	Lapdog media, It is a socialist political party which is inclined towards Muslims	NONE - OFFN
Famous Quotes	Old lions in the wild lay down and die with dignity when they can't hunt anymore.	This quote is used for supreme political leader of BJP.	HATE - PRFN

different classes. The lexical features like satire, slogans, idioms, quotes and code-mixed data are adding to the factor due to which system is producing error. Therefore, a more propped corpus with a substantial learning sample size for each class could give better results in these incidents.

References

[1] R. Kumar, A. K. Ojha, Kmi-panlingua at HASOC 2019: SVM vs BERT for hate speech and offensive content detection, in: P. Mehta, P. Rosso, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019, volume 2517 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 285–292. URL: http://ceur-ws.org/Vol-2517/T3-14.pdf.

[2] 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 Annual Conference of

the North American Chapter of the Association for Computational Linguistics: Human Language Technology (NAACL-HLT), 2019.

[3] T. Ranasinghe, M. Zampieri, An evaluation of multilingual offensive language identification methods for the languages of india, Information 12 (2021). URL: https://www.mdpi.com/2078-2489/12/8/306. doi:10.3390/info12080306.

[4] S. Malmasi, M. Zampieri, Challenges in discriminating profanity from hate speech, Journal of Experimental & Theoretical Artificial Intelligence 30 (2018) 1 – 16.

[5] Z. Waseem, T. Davidson, D. Warmsley, I. Weber, Understanding abuse: A typology of abusive language detection subtasks, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, 2017, pp. 78–84. URL: http://aclweb.org/anthology/W17-3012.

[6] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of ICWSM, 2017.

[7] R. Kumar, A. N. Reganti, A. Bhatia, T. Maheshwari, Aggression-annotated corpus of hindi-english code-mixed data, in: N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Paris, France, 2018.

[8] S. Bhattacharya, S. Singh, R. Kumar, A. Bansal, A. Bhagat, Y. Dawer, b. lahiri, A. K. Ojha, Developing a multilingual annotated corpus of misogyny and aggression, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 158–168.

[9] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Benchmarking aggression identification in social media, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1–11. URL: https://aclanthology.org/W18-4401.

[10] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Evaluating aggression identification in social media, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 1–5.

[11] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval), in: Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval), 2019.

[12] T. Mandl, S. Modha, 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 annual meeting of the Forum for Information Retrieval Evaluation, 2019.

[13] T. Mandl, S. Modha, G. K. Shahi, A. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages, in: FIRE 2020: Forum for Information Retrieval Evaluation, Virtual Event, 16th-20th December 2020, ACM, 2020.

[14] T. Mandl, S. Modha, M. AnandKumar, B. R. Chakravarthi, Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi,

english and german (2020).
[15] 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: FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, 13th-17th December 2021, ACM, 2021.
[16] T. Mandl, S. Modha, G. K. Shahi, H. Madhu, S. Satapara, P. Majumder, J. Schäfer, T. Ranasinghe, M. Zampieri, D. Nandini, A. K. Jaiswal, Overview of the HASOC subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021. URL: http://ceur-ws.org/.
[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[18] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp. 108–122.
[19] M. A. Hearst, Support vector machines, IEEE Intelligent Systems 13 (1998) 18–28. URL: http://dx.doi.org/10.1109/5254.708428. doi:10.1109/5254.708428.