

Confirming the Effectiveness of a Simple Language-Agnostic Yet Very Strong System for Hate Speech and Offensive Content Identification

Yves Bestgen¹

¹*Laboratoire d'analyse statistique des textes - Statistical Analysis of Text Laboratory (LAST - SATLab), Université catholique de Louvain, 10 place Cardinal Mercier, Louvain-la-Neuve, 1348, Belgium*

Abstract

At the 2021 edition of HASOC, the SATLab team proposed a very simple language-agnostic system for hate speech and offensive content identification. This system proved to be extremely effective for the two less resourced languages (e.g., Hindi and Marathi). The present paper describes the use of the same system for task 3 of the 2022 edition of HASOC on hate speech and offensive content identification in Marathi. It consists of a logistic regression applied to character n-grams. It ranked fifth on subtask 3A (macro-F1 = 0.937), quite close to the first ones, second on subtask 3B (macro-F1 = 0.915), very close to the first one, and first (macro-F1 = 0.961) with more than 16 Macro F1 points ahead of the second one in subtask 3C. These results confirm the effectiveness of the approach and suggest that studies evaluating different systems for this kind of problem should employ a character n-gram based approach as a baseline. They also show that the task is extremely simple since all macro-F1s are greater than or equal to 0.915.

Keywords

Character n-grams, logistic regression, low-resource languages

1. Introduction

Hate speech and offensive content on internet is a crucial problem. Insulting or obscene content can hurt many people, but also denigrate entire communities. It is therefore important that web players like Twitter or Facebook are able to identify such content quickly and efficiently. Only the development of automatic detection systems can achieve this. This is the objective of the HASOC evaluation campaigns "Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages" [1, 2].

During the 2021 edition of HASOC, the SATLab proposed a very simple language-agnostic system for hate speech and offensive content identification [3]. This system proved to be extremely efficient for the two less resourced languages of that challenge, ending seventh for Hindi task 1, second for Hindi task 2, and fourth for Marathi.

HASOC 2022 [4] proposes to extend this research path by proposing three subtasks for Marathi. It seemed interesting to determine if the approach proposed last year by the SATLab

Forum for Information Retrieval Evaluation, December 9-13, 2022, India

✉ yves.bestgen@uclouvain.be (Y. Bestgen)

🌐 <https://perso.uclouvain.be/yves.bestgen> (Y. Bestgen)

🆔 0000-0001-7407-7797 (Y. Bestgen)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

was as effective for this year's task.

The following sections of this paper present the three subtasks and the datasets made available for this shared task, the system developed, and the results obtained, which confirms that the proposed approach is a very strong language-agnostic system for hate speech and offensive content identification.

2. Task and material

The task 3 of the HASOC 2022 shared task, to which the SATLab participated, consists of three subtasks in Marathi [5]. The first subtask (3A) requires discriminating between offensive (OFF) and non-offensive (NOT) tweets. The second subtask (3B) requires predicting the type of offense as Targeted (TIN) and Untargeted (UNT) insult. Targeted insult tweets explicitly target an individual, a group or anything else (without any specification). The third subtask (3C) focuses on the target of offences, asking to discriminate between individual (IND), group (GRP) or other (OTH).

It should be noted that the material provided by the organizers is identical for all three subtasks. As a result, only tweets categorized as offensive in the gold label of subtask 3A are categorized in subtask 3B and only tweets categorized as Targeted in the gold label of subtask 3B are categorized according to target type in subtask 3C. The organizers consider subtasks 3B and 3C to be two and three class classification problems, probably because that is the basis on which they evaluate performance. But, for the participants, subtask 3B is a three class classification problem because, when they predict that an offensive instance is a Targeted (TIN) or an Untargeted (UNT) insult, they have to do so on the complete material used for subtask 3A. They therefore do not know which instances are offensive and which are not. Since the systems are penalized if they assign one of the offensive labels to a "neutral" instance, it is indeed a three-category task. Similarly, subtask 3C is a four-class classification problem for the participants because, when they predict that a Targeted (TIN) instance is focused on an individual (IND), on a group (GRP) or on something else (OTH), they have to do so on the complete material used for subtask 3A. They therefore do not know which instances are offensive and targeted and which are not and are penalized if they get it wrong.

No information about the measure of effectiveness was provided to participants during the learning phase. During the testing phase, it appeared that it was Macro-F1, but how this score was calculated was unknown. For these reasons, I do not report here the results on the learning set.

In total, the organizers provided 3103 instances for learning and 510 for the test phase. Table 1 shows the distribution of the learning material in the different categories after deletions of three problematic instances¹. The distribution in the test material is not yet officially known, but a hypothesis is proposed in the result section.

¹Id 1865 contains no text, but is nevertheless offensive, Id 1981 is offensive, but has no label for subtask 2B, and Id 2324 is targeted, but has no label for subtask 3C

Table 1

Dataset statistics on the learning set for the three subtasks. Note: IRR = Irrelevant for that task, but present in the material and should be identified as such.

Subtask 3A		Subtask 3B		Subtask 3C	
NOT	2034	IRR	2034		
		UNT	327	IRR	2361
OFF	1066				
		TIN	739	IND	502
				GRP	157
				OTH	80

3. Proposed system

The proposed system is a very simplified version of the one used for HASOC 2021 and the VarDial challenges [3, 6]. Its features are the following:

- It is only based on character n-grams observed at least twice in the material,
- The n-grams were one to five characters in length,
- The n-grams that start or end a tweet were marked as such,
- Their frequency in the tweet was weighted by means of BM25 ([7, 8]),
- The feature scores for each instance were normalized by the classical L2 regularization.

The parameters, such as the length of the character n-grams, were not set on the HASOC 2022 learning material, but directly taken from the system used for HASOC 2021. These features were provided to the L1-regularized logistic regression from the LIBLinear package [9], with the $-B1$ (bias) option, an approach very simple to use because it only requires the optimization of the regularization parameter C and of the $-wi$ parameters which allow to adjust this C parameter for the different categories.

The optimization of the parameters was performed independently for each subtask by means of an ANSI C program using several successive random grid searches in a 4-fold stratified cross-validation procedure [10]. This produced the parameter values given below. Thus, a different and independent model is build for each subtask.

- Subtask 3A: $c=4.5$, $-w(\text{NOT})=1$, $-w(\text{OFF})=2.1$.
- Subtask 3B: $c=6$, $-w(\text{IRR}^2)=1$, $-w(\text{TIN})=4$, $-w(\text{UNT})=15$.
- Subtask 3C: $c=46$, $-w(\text{IRR})=1$, $-w(\text{IND})=8.04$, $-w(\text{GRP})=74$, $-w(\text{OTH})=115$.

As can be seen, the values of these parameters are very diverse, raising concerns about an overfit problem when applying them to the test material. In particular, the $-wi$ are strongly influenced by the imbalance of the data in the different categories. However, there was no

²IRR = Irrelevant for that task, but present in the material and should be identified as such.

Table 2

Macro-F1 on the test set for the three subtasks. Note: R = Rank.

Task	R	Team	Score	R	Team	Score	R	Team	Score
3A	1	ssncse_nlp	0.975	2	optimize prime	0.959	5	SATLab	0.937
3B	1	hate-busters	0.921	2	SATLab	0.915	3	ssncse_nlp	0.696
3C	1	SATLab	0.961	2	ssncse_nlp	0.793	3	ml_ai_iiitranchi	0.742

reason to believe that the same distribution would be found in the test material. For this reason, I tried another approach, based on a data augmentation procedure in which synthetic data are added in the less populated categories. It has been shown that such an approach can reduce the difficulties encountered when analyzing highly unbalanced data [11]. In the present challenge, this approach has been shown to be slightly beneficial on the cross-validation training material, but useless and even inefficient on the test material. It will therefore not be described in detail here.

4. Official results

The submission site was excellent, as it was last year, but significant problems immediately arose in the scoring procedure. For this reason, I only made two submissions for each subtask out of the five allowed. More than a week after the end of the challenge, the Macro-F1s for subtask 3A were modified. It is these modified scores that are shown in Table 2. This table also shows two benchmarks for each subtask.

The performance of the system is excellent overall, confirming that a simple language-agnostic approach can be very effective in identifying offensive content and in determining whether it is targeted and what types of targets it is aimed at. Another interpretation of this excellent performance is that the tasks proposed by the organizers is extremely simple. Since I have no knowledge of Marathi, it was not possible for me to analyze the most effective features to try to understand the origin of the effectiveness of the proposed system.

A final observation worth mentioning is that the SATLab model predictions produce long runs of identical labels when the instances are ordered by their official Id, without affecting its effectiveness (macro-F1 > 0.91). Figure 1 illustrates this phenomenon. The predictions for the test material are ordered from Id 0 to Id 509 in rows of 50 instances and 10 for the last row. The five categories are distinguished by colors as shown in the legend. To create this graph, I started with the predictions for subtask 3C (macro-F1 = 0.961), which were supplemented with the predictions for subtask 3B when no category had been assigned to an instance for subtask 3C. Since the proposed system is extremely efficient, one can assume that almost all of these labels are correct and therefore concludes that the labels in the test material were not randomly distributed.

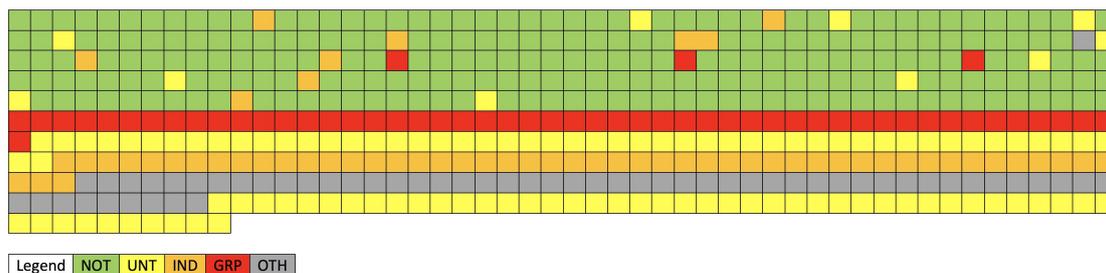


Figure 1: Long runs of identical categories in SATLab model predictions when the instances are ordered by their official Id

5. Conclusion

In the 2021 edition of HASOC, SATLab ended seventh for Hindi task 1, second for Hindi task 2, and fourth for Marathi. This year, a simplified version of that system ended fifth for subtask 3A, second for subtask 3B and first, with a huge margin, for subtask 3C, all in Marathi. These results thus confirm that a very simple language-agnostic approach, based only on character n-grams and logistic regression, can be extremely efficient (Macro-F1 > 0.96 for a four-category classification problem) when the objective is the identification of hate speech and offensive content in Indo-Aryan languages. These results also suggest that studies evaluating different approaches to this kind of problem such as [12] should use a character n-gram based approach as a baseline. It should be noted, however, that this approach, which is also very effective in detecting hyperpartisan news articles [13], is much less effective in identifying passages of text that contain patronizing and condescending language [14].

Acknowledgments

The author is a Research Associate of the Fonds de la Recherche Scientifique - FNRS (Fédération Wallonie Bruxelles de Belgique).

References

- [1] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandalia, A. Patel, Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in indo-european languages, in: P. Majumder, M. Mitra, S. Gangopadhyay, P. Mehta (Eds.), FIRE '19: Forum for Information Retrieval Evaluation, Kolkata, India, December, 2019, ACM, 2019, pp. 14–17. URL: <https://doi.org/10.1145/3368567.3368584>. doi:10.1145/3368567.3368584.
- [2] 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.

- [3] Y. Bestgen, A simple language-agnostic yet strong baseline system for hate speech and offensive content identification, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR Workshop Proceedings, CEUR-WS.org, 2021, pp. 1–10.
- [4] S. Satapara, P. Majumder, T. Mandl, S. Modha, H. Madhu, T. Ranasinghe, M. Zampieri, K. North, D. Premasiri, Overview of the HASOC Subtrack at FIRE 2022: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, in: FIRE 2022: Forum for Information Retrieval Evaluation, Virtual Event, 9th-13th December 2022, ACM, 2022.
- [5] T. Ranasinghe, K. North, D. Premasiri, M. Zampieri, Overview of the HASOC subtrack at FIRE 2022: Offensive Language Identification in Marathi, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, CEUR, 2022.
- [6] Y. Bestgen, Optimizing a supervised classifier for a difficult language identification problem., in: Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), 2021, pp. 96–101.
- [7] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, *Foundations and Trends in Information Retrieval* 3 (2009) 333–389.
- [8] Y. Bestgen, Improving the character ngram model for the DSL task with BM25 weighting and less frequently used feature sets, in: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), Valencia, Spain, 2017, pp. 115–123.
- [9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: A library for large linear classification, *Journal of Machine Learning Research* 9 (2008) 1871–1874.
- [10] Y. Bestgen, LAST at CMCL 2021 shared task: Predicting gaze data during reading with a gradient boosting decision tree approach, in: Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, Association for Computational Linguistics, Online, 2021, pp. 90–96.
- [11] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research* 16 (2002) 321–357.
- [12] M. Zampieri, T. Ranasinghe, M. Chaudhari, S. Gaikwad, P. Krishna, M. Nene, S. Paygude, Predicting the type and target of offensive social media posts in marathi, *Social Network Analysis and Mining* 12 (2022) 77. URL: <https://doi.org/10.1007/s13278-022-00906-8>. doi:10.1007/s13278-022-00906-8.
- [13] Y. Bestgen, Tintin at SemEval-2019 task 4: Detecting hyperpartisan news article with only simple tokens, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 1062–1066. URL: <https://aclanthology.org/S19-2186>. doi:10.18653/v1/S19-2186.
- [14] Y. Bestgen, SATLab at SemEval-2022 task 4: Trying to detect patronizing and condescending language with only character and word n-grams, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 490–495. URL: <https://aclanthology.org/2022.semeval-1.67>. doi:10.18653/v1/2022.semeval-1.67.