

Using Only Character Ngrams for Hate Speech and Offensive Content Identification in Five Low-Ressource Languages

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

This paper describes the system proposed by the SATLab for hate speech and offensive content identification in five low-ressource languages. This language-agnostic system applies a classical supervised learning to character n-grams, using no other data than the learning materials. After optimizing a series of parameters, it ranked first in the Bodo task and second in the Gujarati task, for which the learning material contained only 200 tweets. It also performed well in the Sinhala and Assamese task, but was outperformed by several systems in the Bengali task.

Keywords

Character ngrams, logistic regression, gradient boosting decision tree, low-resource languages

1. Introduction

This year, the SATLab team took part in five tasks proposed by HASOC 2023, the fifth edition of the challenge on Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages [1]. Identifying offensive content on the Internet is both a crucial task and a particularly complex one. It's a very important task because an increasingly large proportion of humanity is informed via the Internet, and because these same people have (a priori) the possibility of disseminating any content they wish. It is therefore very easy to disseminate hateful and offensive content that could harm or affect a large number of users. The sheer volume of content disseminated makes monitoring difficult, especially in languages with limited linguistic resources. HASOC aims to promote the development of automatic techniques for such resource-poor languages [2].

In this NLP field as in many other NLP domains, deep learning and pre-computed embeddings are the preferred solutions, even in low-resource languages [2, 3]. Despite this, the SATLab presented at the two previous HASOC editions a language-agnostic system using only character ngrams as features, with no other linguistic resources [4, 5]. This approach has achieved excellent results, particularly for languages with few linguistic resources. As HASOC 2023 is dedicated to this type of language, the same system has been proposed.

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

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

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

🆔 0000-0001-7407-7797 (Y. Bestgen)



© 2023 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)

A first important feature of this 2023 edition of HASOC is that it includes languages that have never been the subject of this type of challenge, such as Sinhala, Bengali, Bodo and Assamese. The proposed system achieved excellent results for Bodo, finishing first, and for Sinhala, fourth, but very close to the best teams. A second important feature of HASOC 2023 is that for one of the languages, Gujarati, the learning material contained just 200 tweets. The organizers expected participants to explore various techniques to improve the system in a few short settings. The system developed by the SATLab does not use any information sources other than the learning material. Nevertheless, it achieved an excellent second place for this subtask, with a Macro-F1 of 0.8383, 0.0105 points behind the best team.

The remaining of this paper presents the five tasks in which the SATLab took part and the challenge rules. Next, the general characteristics of the proposed systems are described. Finally, the results obtained are discussed.

2. Tasks and Challenge Rules

All the tasks involved identifying hate and offensive content in short messages posted on the Internet, such as tweets or YouTube comments. For each of them, the system was asked to distinguish between messages that included offensive language such as insulting, hurtful, derogatory, or obscene content (HOF) and messages that did not (NOT). All the task languages in which the SATLab participated were poorly endowed with linguistic resources.

In Task 1A ([6]), the tweets were written in Sinhala, an official language of Sri Lanka spoken by just under 20 million people [7], while Task 1B focused on Gujarati, an official language of India spoken by approximately 50 million people. The learning material for Task 1A consisted of 7,500 instances and the test material of 2,500 instances. For Task 1B, there were only 200 tweets for learning and 1,196 for testing.

Task 4 [8] involved three Indian languages and thus three subtasks: Assamese, Bengali [9] and Bodo. The learning material for Assamese consisted of 4036 instances and the test material of 1009 instances. For Bengali, there were 1281 instances for learning and 320 for testing. For Bodo, there were 1679 instances for learning and 420 for testing.

During the test phase of the challenge, five runs could be submitted for tasks 1A and 1B, while for three subtasks of Task 4, five runs could be submitted every day for more than a fortnight. The measure used to evaluate the systems is the Macro-F1 score.

3. Systems

The systems proposed for the five tasks are all derived from the system that achieved excellent results in the 2021 and 2022 editions of HASOC. These were supervised approaches based only on the learning materials provided by the task organizers. Two supervised procedures were used: the LIBLinear L2-regularized logistic regression model (dual, -s 7) for classification [10] and a LightGBM gradient boosting decision tree approach [11]. Since the only features used to categorize instances are ngrams of characters, this approach can be used to analyze any language, including the five in this challenge. This approach is very simple to deploy, as it requires no language-specific resources. It is also powerful because a series of parameters can

be optimized by cross-validation on the learning material. The remainder of this section first presents the parameters affecting feature extraction and then those affecting the supervised learning procedures.

3.1. Feature Extraction

Of the many parameters evaluated for the 2021 and 2022 editions of HASOC, the following two were retained:

- The maximum length of ngrams, which could vary from 4 to 7. In all cases, all ngrams shorter than this maximum value were used.
- The weighting applied to the frequency of each feature in an instance: the sublinear Tfidf and BM25 [12].

In all the systems developed, the minimum frequency of a feature in the material analyzed has been set at 2, and the weighting scheme applied to all features in an instance is the L2 normalization.

3.2. Learning procedures

For the LIBLinear L2-regularized logistic regression model (dual, -s 7), three parameters were evaluated:

- The regularization parameter C.
- The -w1 options for adjusting the parameter C of the HOF category.
- The bias parameter (-B), which shifts the separating hyperplane from the origin.

LightGBM's parameters are far too numerous to present here. They have been optimized by the automatic procedure described in Bestgen [13].

3.3. System Optimization

The parameters presented above were first optimized on the training material using a 4-fold cross-validation procedure stratified by category. Secondly, some trials allowed by the challenge rules were used to try to optimize these parameters for the test set. For each task, both the LIBLinear and LightGBM procedures were evaluated. In the challenge submissions, the whole training material was used.

4. Results

This section successively presents the results of the systems submitted for each of the five tasks.

Table 1

Macro-F1s for the best teams in Task 1A and 1B

1A: Sinhala			1B: Gujarati		
Rank	Team	Macro-F1	Rank	Team	Macro-F1
1	FiRC-NLP	0.8382	1	FiRC-NLP	0.8488
2	Krispy Mango	0.8371	2	SATLab	0.8383
3	AiAlchemists	0.8355	3	Krispy Mango	0.7956
4	SATLab	0.8351	4	AiAlchemists	0.7926
5	Z-AGI Labs	0.8349	5	XAG-TUD	0.7799
6	NAVICK	0.8281	6	SSN_CSE_ML_TEAM	0.7732

4.1. Task 1A: Sinhala

The cross-validation procedure on the training material led to the choice of the following parameters for feature extraction: maximum length of 5 characters and sublinear TfIdf. The cross-validation did not reveal any significant differences between the supervised learning procedures and so both approaches were evaluated on the test material. The system that performed best was an ensemble of three other models: a LIBLinear ($C=8$, $w1=1.8$ and $B=0.2$) and two LightGBM models, the first based on the same features as the LIBLinear and the second based on ngrams ranging from 1 to 7 characters. These three models obtained cross-validation Macro-F1 scores of 0.8018, 0.8289 and 0.8295 respectively. The best of them obtained 0.8304 on the test material. The set of three systems (majority vote) came fourth in the challenge with a Macro-F1 of 0.8351, just 0.0031 behind the best team, as shown in Table 1.

4.2. Task 1B: Gujarati

As a reminder, the training set for this task contained only 200 instances. The cross-validation procedure on this set led to the choice of the following parameters for feature extraction: maximum length of 4 characters and sublinear TfIdf. Cross-validation showed that LightGBM (Macro-F1 = 0.75) was clearly more efficient than LIBlinear (Macro-F1 = 0.68, $C = 3.5$, $B = 1$). This observation was confirmed on the test material (but to a lesser extent) with Macro-F1 values of 0.8188 and 0.8130 respectively.

However, while precision and recall for the LightGBM version were almost identical, LIBLinear’s precision (0.8890) was significantly higher than recall (0.7840), suggesting that the system was assigning too few instances to the HOF category. On the basis of the probabilities of belonging to this category returned by LIBLinear, the proportion of HOFs in the prediction was increased by assigning to this category all instances with a probability greater than or equal to 0.43 (instead of the default value of 0.50), raising the proportion of HOFs in predictions on the test material from 0.20 to 0.28. This simple trick, which increased recall to 0.83 while only reducing precision to 0.85, enabled the system to gain 0.02 points and take second place in the challenge with a Macro-F1 of 0.8383, 0.0105 points behind the best team and more than 0.04 ahead of the third-placed team (see Table 1). It would be interesting to compare the performance of these systems using the bootstrap confidence intervals [14] to determine whether they are of any practical use. When making such a comparison, it will be necessary to take into account

Table 2

Parameters and the Macro-F1s for the three languages in Task 4

Language	Ngram length	Weighting	C	w1	Macro-F1		Rank
					CV	Test	
Assamese	6	BM25	6	0.72	0.689	0.715	4
Bengali	5	TfIdf	2.7	1.75	0.656	0.671	9
Bodo	7	TfIdf	10	1.15	0.817	0.857	1

the resources employed by each system [15].

The proposed approach, which uses only 200 tweets for training, is therefore very effective. However, there is a significant and unexpected difference between the performance on the training material with a maximum Macro-F1 of 0.75 in cross-validation and the performance on the test material with a Macro-F1 of 0.84. This difference may be due to the fact that cross-validation training is carried out on only 150 tweets, whereas the test phase uses 200.

4.3. Task 4: Assamese, Bengali and Bodo

The results for these three subtasks are presented together because the proposed systems are very similar. The LIBLinear procedure is used in each case. Table 2 shows the parameters derived from the cross-validation and the Macro-F1 achieved on using the CV and on the test set. These systems ranked first for Bodo with a 0.006 lead over the 2nd team, fourth for Assamese with a 0.019 difference from the first and ninth for Bengali with a 0.10 difference from the first.

5. Conclusion

The SATLab approach for identifying offensive content in short social network posts proved highly effective for four of the five languages (Bodo, Gujarati, Sinhala and Assamese), but much less so for the last one (Bengali), since the difference with the best team for these two languages is almost 0.10 Macro-F1 score. The origin of these differences is unknown to me. Only a reading of the organizers' synthesis [1] could reveal whether there are differences between these tasks or between the systems presented to perform them.

The efficiency obtained for Gujarati is quite astonishing and unexpected for an approach that employs no other resources than the learning material, which is limited for this language to 200 instances. This result suggests that it would be interesting to repeat all the HASOC tasks proposed over the last five years and determine for each of them the impact of the number of instances available for learning on performance in the test phase. To be honest, I doubt that such good results could be obtained for all of them. The difference in performance in Task 4 between Bodo and the other two languages also merits further analysis.

Acknowledgments

The author wishes to thank the organizers of this shared task for putting together this valuable

event. He is a Research Associate of the Fonds de la Recherche Scientifique - FNRS (Fédération Wallonie Bruxelles de Belgique). Computational resources have been provided by the super-computing facilities of the Université Catholique de Louvain (CISM/UCL) and the Consortium des Equipements de Calcul Intensif en Fédération Wallonie Bruxelles (CECI).

References

- [1] T. Ranasinghe, K. Ghosh, A. S. Pal, A. Senapati, A. E. Dmonte, M. Zampieri, S. Modha, S. Satapara, Overview of the HASOC subtracks at FIRE 2023: Hate speech and offensive content identification in assamese, bengali, bodo, gujarati and sinhala, in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE 2023, Goa, India. December 15-18, 2023, ACM, 2023.
- [2] 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.
- [3] T. Mandl, S. Modha, A. Kumar, 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: P. Majumder, M. Mitra, S. Gangopadhyay, P. Mehta (Eds.), FIRE 2020: Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, ACM, 2020, pp. 29–32. URL: <https://doi.org/10.1145/3441501.3441517>. doi:10.1145/3441501.3441517.
- [4] Y. Bestgen, A simple language-agnostic yet strong baseline system for hate speech and offensive content identification, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR Workshop Proceedings, CEUR-WS.org, 2021, pp. 1–10.
- [5] Y. Bestgen, Confirming the effectiveness of a simple language-agnostic yet very strong system for hate speech and offensive content identification, in: K. Ghosh, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, CEUR Workshop Proceedings, CEUR-WS.org, 2022, pp. 1–6.
- [6] 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.
- [7] T. Ranasinghe, I. Anuradha, D. Premasiri, K. Silva, H. Hettiarachchi, L. Uyangodage, M. Zampieri, Sold: Sinhala offensive language dataset, arXiv preprint arXiv:2212.00851 (2022).
- [8] K. Ghosh, A. Senapati, A. S. Pal, Annihilate Hates (Task 4, HASOC 2023): Hate Speech Detection in Assamese, Bengali, and Bodo languages, 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.

- [9] N. Romim, M. Ahmed, H. Talukder, M. Saiful Islam, Hate speech detection in the Bengali language: A dataset and its baseline evaluation, in: M. S. Uddin, J. C. Bansal (Eds.), *Proceedings of International Joint Conference on Advances in Computational Intelligence*, Springer Singapore, Singapore, 2021, pp. 457–468.
- [10] 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.
- [11] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A highly efficient gradient boosting decision tree, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems* 30, Curran Associates, Inc., 2017, pp. 3146–3154. URL: <http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf>.
- [12] 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.
- [13] 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. URL: <https://aclanthology.org/2021.cmcl-1.10>. doi:10.18653/v1/2021.cmcl-1.10.
- [14] Y. Bestgen, Please, don't forget the difference and the confidence interval when seeking for the state-of-the-art status, in: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 2022, pp. 5956–5962. URL: <https://aclanthology.org/2022.lrec-1.640>.
- [15] J. Dodge, S. Gururangan, D. Card, R. Schwartz, N. A. Smith, Show your work: Improved reporting of experimental results, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2185–2194. URL: <https://www.aclweb.org/anthology/D19-1224>. doi:10.18653/v1/D19-1224.