Hate Speech Detection Using Transformer Ensembles on the HASOC Dataset

Pedro Alonso , Rajkumar Saini , Gÿorgy Kov́acs
Embedded Internet Systems Lab, Lule̊a University of Technology, Lule̊a, Sweden
MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary

Year of publication: September 2020
***N**umber of citations: 17*

Overview

1. Introduction
2. Related work
3. Dataset
4. Contribution
5. Results
6. Conclusion

Introduction

What is hate speech?

We define hate speech as a direct attack on people based on what we call protected characteristics—race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We also provide some protections for immigration status. We define attack as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation. [1]

- Facebook –

Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. [2]

- Twitter -

"We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, and veteran Status"[3]

- YouTube -

[1] "Community standards." [Online]. Available: https://www.facebook.com/ communitystandards/objectionable_content/

[2] "twitters policy on hate help." [Online]. Available: https://archive.org/details/perma_cc_2XYS-VWJB

[3] "Hate speech policy - ful conduct | twitter youtube help." [Online]. Available: https://support.google. com/youtube/answer/2801939?hl$=$en

Introduction cont.

- Growth of social-media usage, raises a platform to a new kind of social dilemma namely cyberbullying
- HS detection is still a challenge for the research community and policy makers as humans find loopholes to trick those algorithms [5]
- Challenge of detecting hate speech within online user communication due to its vast scope and the complexity
- Secure the freedom of speech [6]

[5] A. Oboler, "Solving antisemitic hate speech in social media through a global approach to local action," in Volume 5 Confronting Antisemitism in Modern Media, the Legal and Political Worlds. De Gruyter, 2021, pp. 343–368. 2

[6] T. M. Massaro, "Equality and freedom of expression: The hate speech dilemma," Wm. & Mary L. Rev., vol. 32, p. 211, 1990. 4

Related work

- Rule-based system (Smokey), template-based or keyword-based systems
- Kwok and Wang [7]
  - Feature extraction using Bag-of-Words (BoW)
  - Naïve Bayes classifier for the detection of racism against black people on Twitter
- Grevy et al. [8]
  - Feature extraction using Bag-of-Words (BoW)
  - Support Vector Machines (SVMs)
- Deep Leaning approaches

  - RNN, CNN RNN+CNN
- Transformer approaches

  - BERT[9]

  - Ensembles of transformers [10, 11]

Dataset – HOSAC[12]

- Hindi, German and English
- Twitter archive and pre-classified by a machine learning system
- HASOC has two sub-task for all three languages:
  - Task A : binary classification problem (Hate and Not Offensive)
  - Task B : fine-grained classification problem for three classes (HATE) Hate speech, OFFENSIVE and PROFANITY

Tweet	Label
@piersmorgan Dont watch it then. #dickhead	NOT
This is everything. #fucktrump https://t.co/e2C48U3pss	HOF
I stand with him ...He always made us proud 🤦‍♂️#DhoniKeepsTheGlove	NOT
@jemelehill He's a cut up #murderer	HOF
#fucktrump #impeachtrump 😂😂😂😂😂😂😂 @ Houston, Texas https://t.co/8QGgbWtOAf	NOT

Dataset – HOSAC[12]

- HASOC data (English language data) to be tackled using our methods.
- 6712 tweets (the training and test set containing 5852 and 860 tweets, respectively)
- Annotated into the following categories:
  - NOT: tweets not considered to contain hateful or offensive content
  - HOF: tweets considered to be hateful, offensive, or profane

Tweet	Label
@piersmorgan Dont watch it then. #dickhead	NOT
This is everything. #fucktrump https://t.co/e2C48U3pss	HOF
I stand with him ...He always made us proud 🙌#DhoniKeepsTheGlove	NOT
@jemelehill He's a cut up #murderer	HOF
#fucktrump #impeachtrump 😂😂😂😂😂😂😂 @ Houston, Texas https://t.co/8QGgbWtOAf	NOT

Dataset – HOSAC – OffensEval[13]

Tweet	Score
@USER And cut a commercial for his campaign.	0.2387
@USER Trump is a fucking idiot his dementia is getting worse	0.8759
Golden rubbers in these denim pockets	0.3393
Hot girl summer is the shit!!! #period	0.8993

Contribution

- **Data preprocessing**
- Lacks proper grammar/punctuation and contains many paralinguistic elements

  e.g. URLs, emoti cons, emojis, hashtags
- Consecutive white space characters were replaced by one instance
- Extra white space characters were added between words and punctuation marks
- @-mentions and links were replaced by the character series @USER and URL
- All emojis and emoticons were removed
- Removed hashtag characters

Contribution

- A variant of BERT : **RoBERTa**[14]
- 5-fold ensemble training method using the RoBERTA model
  - Split the HASOC train set into five equal parts, each consisting of 1170 tweets
  - Balanced dataset in each part
  - Created a training set using the remaining tweets from the original training set, for each development set
  - Used each fold to train separate RoBERTa models
  - Final model → ensemble of the five individual models

Results - Model performance

Model	Fold	$HASOC_{only}$	$HASOC_{OffensEval}$
Macro F_1-score	1st	0.7586	0.7964
	2nd	0.7681	0.7855
	3rd	0.7688	0.7943
	4th	0.7524	0.7929
	5th	0.7758	**0.8029**
	Ensemble	**0.7945**	0.7976
Weighted F_1-score	1st	0.8125	0.8507
	2nd	0.8165	0.8402
	3rd	0.8244	0.8474
	4th	0.8415	0.8485
	5th	0.8327	**0.8537**
	Ensemble	**0.8426**	0.8504

Alonso, P., Saini, R. and Kovács, G., 2020, October. Hate speech detection using transformer ensembles on the hasoc dataset. In *International conference on speech and computer* (pp. 13-21). Springer, Cham.

[15] Alonso, P., Saini, R. and Kovács, G., 2020, October. Hate speech detection using transformer ensembles on the hasoc dataset. In *International conference on speech and computer* (pp. 13-21). Springer, Cham.

Conclusion

Outperforming the best performing system in the literature attaining a weighted F1- score of 0.8426.

Further improve by leveraging more training data, achieving a weighted F1-score of 0.8504.

Reference

[1] “Community standards.” [Online]. Available: https://www.facebook.com/ communitystandards/objectionable_content/

[2] “twitters policy on hate help.” [Online]. Available: https://archive.org/details/perma_cc_2XYS-VWJB

[3] “Hate speech policy - ful conduct | twitter youtube help.” [Online]. Available: https://support.google.com/youtube/answer/2801939?hl$=$en

[4] AlKhamissi, B., Ladhak, F., Iyer, S., Stoyanov, V., Kozareva, Z., Li, X., Fung, P., Mathias, L., Celikyilmaz, A. and Diab, M., 2022. ToKen: Task Decomposition and Knowledge Infusion for Few-Shot Hate Speech Detection. arXiv preprint arXiv:2205.12495.

[5] A. Oboler, “Solving antisemitic hate speech in social media through a global approach to local action,” in Volume 5 Confronting Antisemitism in Modern Media, the Legal and Political Worlds. De Gruyter, 2021, pp. 343–368. 2

[6] T. M. Massaro, “Equality and freedom of expression: The hate speech dilemma,” Wm. & Mary L. Rev., vol. 32, p. 211, 1990. 4

[7] Kwok, I., Wang, Y.: Locate the hate: Detecting tweets against blacks. In: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence. p. 1621–1622. AAAI’13, AAAI Press (2013)

Reference

[8] Greevy, E., Smeaton, A.F.: Classifying racist texts using a support vector machine. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 468–469. SIGIR '04, Association for Computing Machinery, New York, NY, USA (2004). https://doi.org/10.1145/1008992.1009074

[9] Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune bert for text classification? In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds.) Chinese Computational Linguistics. pp. 194–206. Springer International Publishing, Cham (2019)

[10] Nina-Alcocer, V.: Vito at HASOC 2019: Detecting hate speech and offensive content through ensembles. In: Mehta, P., Rosso, P., Majumder, P., Mitra, M. (eds.) Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019. CEUR Workshop Proceedings, vol. 2517, pp. 214–220. CEUR-WS.org (2019)

[11] Nourbakhsh, A., Vermeer, F., Wiltvank, G., van der Goot, R.: sthruggle at SemEval-2019 task 5: An ensemble approach to hate speech detection. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 484–488. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/S19-2086

[12] Mandl, T., Modha, S., Mandlia, C., Patel, D., Patel, A., Dave, M.: HASOC - Hate Speech and Offensive Content identification in indo-european languages, https://hasoc2019.github.io, accessed on 2019-09-20

[13] Z ampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., C¸ oltekin, c.: SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In: Proceedings of SemEval (2020)

[14] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)

[15] Alonso, P., Saini, R. and Kovács, G., 2020, October. Hate speech detection using transformer ensembles on the hasoc dataset. In *International conference on speech and computer* (pp. 13-21). Springer, Cham.

Thank you!

Any Questions?

You can find me at:
dinuja.21@cse.mrt.ac.lk