Are Hate Speech Classification Results Reproducible? An Approach Using Deep Learning

Paula Fortuna[1,2] Juan Soler-Company[2] Sérgio Nunes[1,3]
(1) INESC TEC, (2) Pompeu Fabra University and (3) FEUP, Porto University

1. Hate speech? #%&#!

This work appears in the context of the Stop PropagHate project, that aims at fighting online hate speech.

> Hate speech is language that **incites hate** against **groups**, based on **specific characteristics** and it can occur with **different linguistic styles**, even in subtle forms or when humour is used [1].

2. Study goal

In this work, we have the goal to reproduce a state-of-the-art hate speech classifier, in order to find a classifier with good performance.

We choose to reproduce a specific study [2], because this is a highly cited paper, describing a classifier with unique good performance (F=0.93), of which the authors publish their code.

3. State of the art

Deep Learning techniques are quickly gaining ground in the area. Different studies proved that deep learning algorithms outperform classical Machine Learning approaches (e.g. [2])

4. Methodology

We follow a methodology of 10-fold cross-validation with holdout validation [3] . A scheme of the used classification method [2] is presented in the Figure 1.

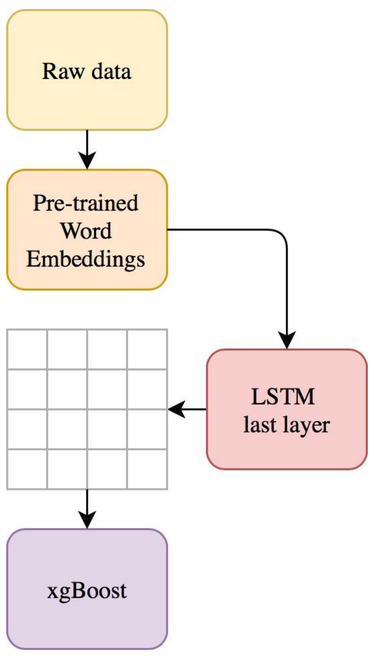


Figure 1 - classifier used in the experiments

5. Replication problem

We tried to replicate the aforementioned study [2] but we faced one main difficulty: because of a bug the effect of using 10-fold cross-validation was eliminated. As a proof of this problem, we can see that the successive values of F1 score found in the 10 iterations increases (Figure 2).

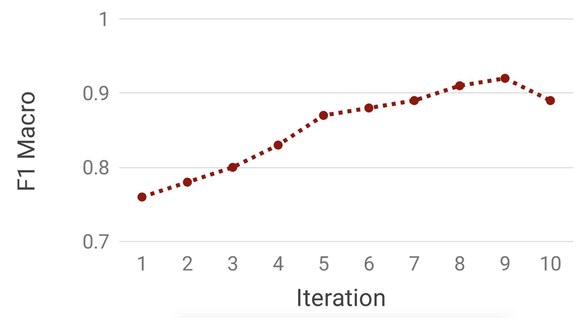


Figure 2 - Cross validation results

6. Experiment & results

We conducted 3 classification experiments using a standard dataset [4] , the HatEval [5] and the OffensEval [6] datasets. We compare these with the results from the original study [2] (Figure 3).

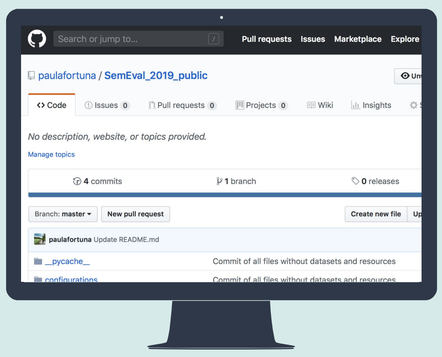
Figure 3 - Experiments comparison

7. Conclusion

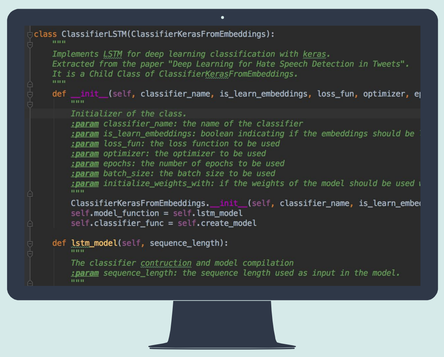
- Despite the good results we achieved, we could not replicate the targeted study.
- This allowed us to see the importance of properly sharing code in this field.
- We took our results into account and developed this work in a reproducible way.

Project develoment towards reproducibility

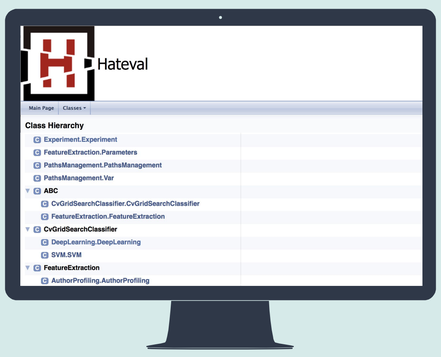
Shared code in GitHub



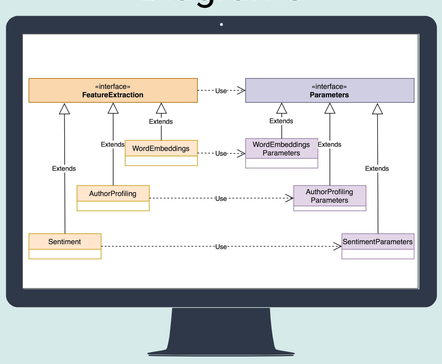
Thoroughly commented code



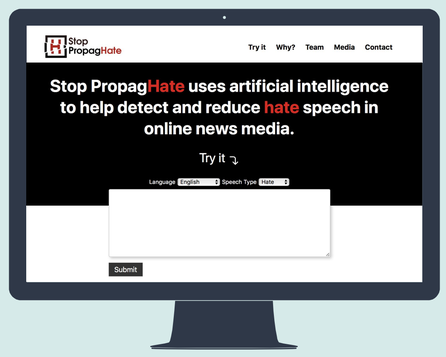
Doxygen



Class Diagrams



User interface



Acknowledgements

This work was partially funded by the Google DNI grant Stop PropagHate. The printing of the poster was kindly supported by the EITIC.

Stop PropagHate | Google DIGITAL NEWS INITIATIVE | INESCTEC | U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO | upf. Universitat Pompeu Fabra Barcelona — Engineering and Information and Communication Technologies Academic Coordination Unit

References

[1] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4):85.

[2] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, pages 759–760. International World Wide Web Conferences Steering Committee.

[3] Francois Chollet. 2017. Deep learning with python. Manning Publications Co.

[4] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter.In Proceedings of NAACL-HLT, pages 88–93

[5] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019). Association for Computational Linguistics.

[6] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval).