|  |
| --- |
| **A Comparative study of identifying hate and offensive comments on social media platforms using text classification models and NLP techniques**    - By Eranki Vasistha |
|  |
| **Final Project Report** |
|  |
|  |
|  |

Abstract

In today's internet-dependent world, social media platforms like Twitter, Instagram, and Facebook have interconnected the globe in a truly fascinating way, enabling instant communication between individuals across geographical boundaries, whether in the United States, Japan, India, or even the most remote locations. However, this increased connectivity has also led to a surge in online cyberbullying and the proliferation of inappropriate racial, sexist, and abusive slurs on these platforms. All these hateful comments are posted on social media in text form. Through this project, I aim to develop an automated identifier that detects such hate comments, which would help us in enabling the blocking of offending users or deletion of harmful content, or both

Brief overview of tasks performed

This project focuses on detecting hate speech on social media using machine learning models. The aim is to evaluate and compare the performance of various algorithms in classifying text into three categories: "Hate Comment," "Offensive Comment," and "Neither." Key tasks included feature selection, data preprocessing, and training multiple models for text classification. Error analysis and result evaluation were conducted for each model, followed by hyperparameter tuning to enhance performance. Finally, the project discusses potential improvements and future directions for this work.

Data Description

The dataset consists of 24,783 labeled text records, with each record containing a tweet and its corresponding label indicating one of three categories: "Hate Comment" (the minority class), "Offensive Comment," or "Neither." The data is split into a training set comprising 17,348 records (70% of the dataset) and a test set of 7,435 records (30%), providing a balanced setup for training and evaluation.

The dataset distribution, visualized in the pie charts in figure -1, highlights the proportion of each label category in both the training and testing datasets. Overall, the dataset consists of 19,190 offensive comments, 1,430 hate comments, and 4,163 "neither" comments. This imbalance is also reflected in the training dataset, where offensive comments dominate with 13,319 instances (76.8%), followed by 2,926 "neither" comments (16.9%) and 1,103 hate comments (6.4%). Similarly, in the testing dataset, offensive comments constitute 79% of the records, while "neither" and hate comments make up 16.6% and 4.4%, respectively. The charts clearly illustrate the skewed distribution toward offensive comments, which will influence model training and evaluation.

Example Datasets

1. MS Word STREAM Tools

This Microsoft Word file was updated in 2016 with STREAM Tools, designed for creating well-formatted reports and papers with Microsoft Word (Mamishev, 2010; Mamishev, 2013).

Limitations

ACL 2023 requires all submissions to have a section titled “Limitations”, for discussing the limitations of the paper as a complement to the discussion of strengths in the main text. This section should occur after the conclusion, but before the references. It will not count towards the page limit. The discussion of limitations is mandatory. Papers without a limitation section will be desk-rejected without review.

While we are open to different types of limitations, just mentioning that a set of results have been shown for English only probably does not reflect what we expect. Mentioning that the method works mostly for languages with limited morphology, like English, is a much better alternative. In addition, limitations such as low scalability to long text, the requirement of large GPU resources, or other things that inspire crucial further investigation are welcome.

Ethics Statement

Scientific work published at ACL 2023 must comply with the ACL Ethics Policy.[[1]](#footnote-1) We encourage all authors to include an explicit ethics statement on the broader impact of the work, or ethical considerations after the conclusion but before the references. The ethics statement will not count toward the page limit (8 pages for long, 4 pages for short papers).

Acknowledgments

This document has been adapted by Jordan Boyd-Graber, Naoaki Okazaki, Anna Rogers from the template for earlier ACL, EMNLP and NAACL proceedings, including those for EACL 2023 by Isabelle Augenstein and Andreas Vlachos and EMNLP 2022 by Yue Zhang, Ryan Cotterell and Lea Frermann.

References

Rie Kubota Ando and Tong Zhang. 2005. A frameworkfor learning predictive structures from multiple tasksand unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. [Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing](http://aclweb.org/anthology/P16-1001). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1–11. <https://doi.org/10.18653/v1/P16-1001>.

Dan Gusfield. 1997. *Algorithms on Strings, Trees**and Sequences*. Cambridge University Press, Cambridge, UK.

Mary Harper. 2014. [Learning from 26 languages: Pro- gram management and science in the babel program](http://aclweb.org/anthology/C14-1001). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, page 1. <http://aclweb.org/anthology/C14-1001>.

Alexander V. Mamishev and Murray Sargent. 2013. *Creating Research and Scientific Documents Using Microsoft Word*. Microsoft Press, Redmond, WA.

Alexander V. Mamishev and Sean D. Williams. 2010. *Technical Writing for Teams: The STREAM Tools Handbook*. Wiley-IEEE Press, Hoboken, NJ.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

1. Appendices

Appendices are added after the References section by restarting the header numbering using style “A, B, C”.

1. <https://www.aclweb.org/portal/content/acl-code-ethics> [↑](#footnote-ref-1)