Online Hate Speech Detection on Twitter

Jainul N. Vaghasia
Department of Computer Science
University of Washington
Seattle, WA 98105
jnv3@cs.washington.edu

1 Proposal

1.1 Datasets

Sexism/Racism dataset (Waseem and Hovy [2016]), Hate dataset (Davidson et al. [2017]), expecting use of synthetic data to explore additional behavior.

1.2 Idea

Determining whether a tweet is offensive or not is an online binary classification problem that is inherently challenged by class imbalance and concept drift—it is expected that offensive tweets appear less often than not, and that the underlying distribution of such tweets varies based on socio-political events (Golbeck et al. [2017]). In this project, we explore an online approach based on selective resampling of a subset of past training data, and apply it to a neural network classifier based on word embeddings for predicting the class, offensive or not, of a tweet. Malialis et al. [2018]'s novel, online algorithm, albeit simple, is highly intuitive and produces high quality results on synthetic data, whereas Kshirsagar et al. [2018]'s work uses relatively fewer parameters than its counterparts on offline model while achieving high F1 scores.

We stress-test the union of both ideas and explore whether it works equally well while addressing the aforementioned practical problems one might encounter with the given classification problem. If this approach works, what does each idea bring to the approach? Do we experience any counteraction on working of one algorithm as a result of the other one and vice-versa? If the approach fails, what makes this problem uniquely difficult to solve: online vs. offline, magnitude of imbalance, type of drift? We wish to explore these potential questions in this project.

1.3 Software

We expect to implement all the required algorithms and model online data streams using basic Python libraries like Numpy, Scipy, etc.

1.4 Expected Progress by Milestone

By the Milestone, we expect to produce experimental evidence on effectiveness of the approach on online class imbalance.

References

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515, 2017.

Preprint. Work in progress.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, pages 229–233, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4896-6. doi: 10.1145/3091478.3091509. URL `http://doi.acm.org/10.1145/3091478.3091509`.

Rohan Kshirsagar, Tyus Cukuvac, Kathleen McKeown, and Susan McGregor. Predictive embeddings for hate speech detection on twitter. *arXiv preprint arXiv:1809.10644*, 2018.

Kleanthis Malialis, Christoforos Panayiotou, and Marios M. Polycarpou. Queue-based resampling for online class imbalance learning. In *ICANN*, 2018.

Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N16-2013`.