Hate Speech Detection on Twitter

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Jainul Vaghasia | Machine Learning CSE546 Autumn 2018 | University of Washington

Abstract

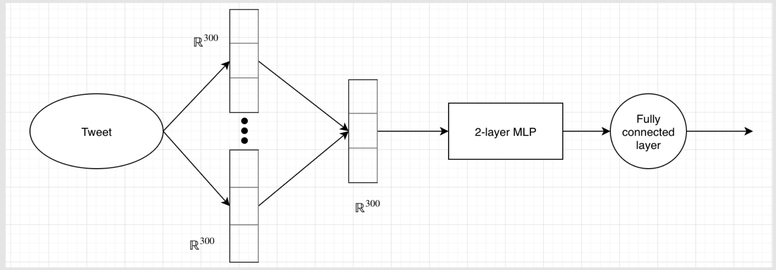
- We explore a neural network based method for classifying the presence of hate speech in a tweet. The network is based on word embeddings and average pooling, and produces state-of-the art results on offline model. The method when combined with resampling of subset of past training data performs well under severe class imbalance in online setting. With correct queue size, we obtain robust behavior against moderate concept drift in the stream as well.

Problem

- Motivation: Offensive behavior on social network platforms has become rather common–73% Twitter users have witnessed harassment online [1].
- Binary classification problem: classify a tweet as offensive or not.
- Issues:
  - What is offensive? Human communication is inherently ambiguous. We adopt the following definition of offensive: "language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group."
  - Class Imbalance: Offensive tweets are almost always less probable than not offensive ones.
  - Concept drift: The phenomenon in which the underlying distribution of classes varies over time. In our case, these might be socio-political events that trigger rash behavior on Twitter.
- In this project we address issues of class imbalance and moderate cases of abrupt concept drifts.
- Model with minimal features which can be adapted easily for different social networks.

Model

- Offline model [2]
  - Based on word embeddings and average pooling of words in the tweets with weighted cross entropy loss.



- Online model [3]
  - Maintains a queue of subset of previous data from both classes—helps handle class imbalance through an analogue of oversampling.
  - Fixed size queues handle concept drift.

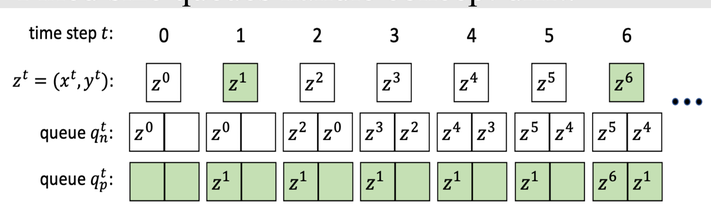


Fig. 1: Example of $Queue_2$ resampling

Offline Results

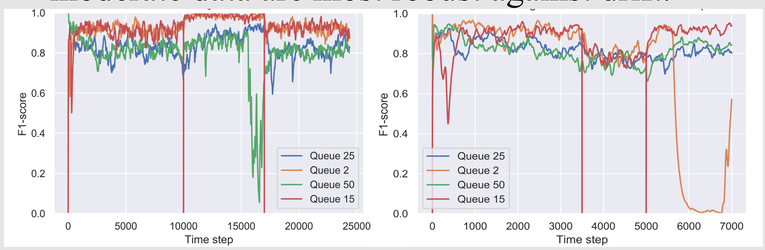
- State-of-the-art F1 results on two datasets with minimal feature engineering and additional data.
- HAR dataset has more subtle hate speech, so the performance is comparatively lesser than that on HATE dataset.

Method	HATE	HAR
Logistic Regression with Feature Engineering	0.93	0.68
Ours	0.94	0.70

Online Results

- Class Imbalance
  - Queue 2 performs the best with Queue 15 eventually catching up.
  - Sensitive to Queue size with excessive oversampling as seen in Queue 25 and Queue 50.

- Concept Drift
  - Model tested on abrupt changes in underlying distribution of tweets (Performance metric is reset to 0 at each drift point).
  - Depending on the direction of the drift, the queues with few or excessive data may not recover quickly from the drift. Queues with moderate data are most robust against drift.



References

1. Golbeck et. al, doi: 10.1145/3091478.3091509
2. Kshirsagar et. al, *arXiv:1809.10644*, 2018.
3. Malialis et. al, Queue-based resampling for online class imbalance learning. In *ICANN, 2018.*