Deep Learning for Hate Speech Detection in Tweets

Last edited by **Hamed Babaei** 1 minute ago

Deep Learning for Hate Speech Detection in Tweets

Abstract:

- They define the task as being able to classify a tweet as `racist, sexist or neither`
- We perform extensive experiments with multiple deep learning architectures to learn semantic word embeddings to handle complexity
- Their experiments on a benchmark dataset show that such deep learning methods outperform state-of-the-art char/word n-gram methods by ∼18 F1points.

Introduction

- On Twitter, hateful tweets are those that contain abusive speech targeting individuals (cyber-bullying, a politician, a celebrity,a product) or particular groups (a country, LGBT, a religion, gender, an organization, etc.).
- Logistic Regression, Random Forest, SVMs, Gradient Boosted Decision Trees (GBDTs) and Deep Neural Networks(DNNs) experimented
- feature spaces for these classifiers are, CNN, LSTM, FastText
- A baseline feature spaces are char n-grams[6] (this is a dataset paper- I added this dataset reference into dataset category!), TFIDF, Bag of Words vectors(BoWV).
- contributions of the paper:
  - (1) deep learning methods investigated
  - (2) various semantic embedding like: char ngrams/ TFIDF, BoWV over GloVe and task specific embedding learned using FastText, CNN, LSTM.
  - (3) beating state-of-the-art methods by a large margin (∼18 F1 points better)

Proposed Approach:

- Baseline Methods: (1) char n-grams (dataset paper) as a state-of-the-art. (2) TFIDF, (3) BoWV which uses the average of the word (GloVe) embedding.
- Proposed Methods: for each following methods either `random embeddings` or `GloVe` embedding used: CNN, LSTM (to capture long dependencies), FastText similar to the BoWV model but allows update of word vector through back-propagation during training. opposite of BoWV. All models fine-tuned.

Dataset: Experimentation done on a dataset of 16K annotated tweets. 3383 are labeled as sexist, 1972 as racist, and the remaining are marked as neither.

Experiments and Results:

- They performed 10-fold cross-validation and reported P, R, F1
- They `adam` used for CNN and LSTM, and `RMS-Prop` for FastText as our optimizer
- They performed training in batches of size 128 for CNN & LSTM and 64 for FastText

Table 1: Comparison of Various Methods (Embedding Size=200 for GloVe as well as for Random Embedding)

	Method	Prec	Recall	F1
Part A: Baselines	Char n-gram+Logistic Regression [6]	0.729	0.778	0.753
	TF-IDF+Balanced SVM	0.816	0.816	0.816
	TF-IDF+GBDT	0.819	0.807	0.813
	BoWV+Balanced SVM	0.791	0.788	0.789
	BoWV+GBDT	0.800	0.802	0.801
Part B: DNNs Only	CNN+Random Embedding	0.813	0.816	0.814
	CNN+GloVe	0.839	0.840	0.839
	FastText+Random Embedding	0.824	0.827	0.825
	FastText+GloVe	0.828	0.831	0.829
	LSTM+Random Embedding	0.805	0.804	0.804
	LSTM+GLoVe	0.807	0.809	0.808
Part C: DNNs + GBDT Classifier	CNN+GloVe+GBDT	0.864	0.864	0.864
	CNN+Random Embedding+GBDT	0.864	0.864	0.864
	FastText+GloVe+GBDT	0.853	0.854	0.853
	FastText+Random Embedding+GBDT	0.886	0.887	0.886
	LSTM+GloVe+GBDT	0.849	0.848	0.848
	LSTM+Random Embedding+GBDT	**0.930**	**0.930**	**0.930**

Table 2: Embeddings learned using DNNs clearly show the "racist" or "sexist" bias for various words.

Target Word	Similar words using GloVe	Similar words using task-specific embeddings learned using DNNs
pakistan	karachi, pakistani, lahore, india, taliban, punjab, islamabad	mohammed, murderer, pedophile, religion, terrorism, islamic, muslim
female	male, woman, females, women, girl, other, artist, girls, only, person	sexist, feminists, feminism, bitch, feminist, blonde, bitches, dumb, equality, models, cunt
muslims	christians, muslim, hindus, jews, terrorists, islam, sikhs, extremists, non-muslims, buddhists	islam, prophet, quran, slave, jews, slavery, pedophile, terrorist, terrorism, hamas, murder

tiple classifiers but report results mostly for GBDTs only, due to lack of space.

As the table shows, our proposed methods in part B are

Author Github Page: https://github.com/pinkeshbadjatiya/twitter-hatespeech

BibeText:

```
@article{Badjatiya_2017,
   title={Deep Learning for Hate Speech Detection in Tweets},
   ISBN={9781450349147},
   url={http://dx.doi.org/10.1145/3041021.3054223},
   DOI={10.1145/3041021.3054223},
   journal={Proceedings of the 26th International Conference on World Wide Web Companion - WW
   publisher={ACM Press},
   author={Badjatiya, Pinkesh and Gupta, Shashank and Gupta, Manish and Varma, Vasudeva},
   year={2017}
```