Detecting Offensive Language in Online Texts

Elif Kilik
Student ID: s0197376

The Goal of This Research

Experiments with Machine Learning and Deep Learning techniques to detect offensive language:

- OffensEval 2019 Sub-task A: Label text as offensive vs. not-offensive
- Comparing traditional machine learning and deep learning approaches in NLP
- Experiments on different test sets:
    - Generalizability of models
    - In-domain and out-domain settings

The Datasets

The Offensive Language Identification Dataset (OLID) from OffensEval 2019 → 14100 annotated tweets (13240 in training and 860 in test sets)

Additional data-sets for testing:

- Reddit (1200 comments)
- Wikipedia (1200 posts)
- Textgain (1276 tweets)

Data-set Name	Offensive	Not-Offensive
OLID Training	4400	8840
OLID Test	240	620
Out-Domain: Reddit	543	664
Out-Domain: Wikipedia	600	600
Out-Domain: Textgain	188	1088

Table 1: Distribution of offensive text in different data sets

Text Cleaning

The quality of pre-processing step is “the key factor in boosting the performance” of NLP models (Caselli et al., 2020)

Cleaning steps:

- Mentions and URL’s replaced with @USER and URL (already done)
- Hashtag signs removed, hashtags represented as words.
- Repetitive use of letters and punctuations normalized.
- Self-censored profanity
- Emojis replaced with words
- For SVM: further normalization → lemmatization and POS tagging with spacy, tokenization with NLTK TweetTokenizer and word_tokenizer

Experiments with SVM

Widely used in text classification tasks (examples: Markov and Daelemans, 2021; Zampieri et al., 2019).

Features included tf-idf weighted n-gram vector representations of:

- Tokenized and lemmatized clean tweets
- POS-tagging
- NRC lexicon for emotion associations
- Insults lexicon from Bassignana et al. (2018)

Optimized with Grid Search. 5-fold Cross-validation showed stable model scores.

Experiments with BERT

- The bert-base-uncased model from Huggingface transformers library was fine-tuned with different parameters.
- Fine-tuning methods are applied with and without pre-training → pre-trained model performed better.
- The final model was chosen based on macro-averaged scores, as well as analysis of training and validation losses.

Hyper-parameters	**Value**
training batch size	32
learning rate	1e-6
warmup steps	0
training epochs	15
adam epsilon	1e-8
max. sequence length	256

Results

- BERT outperforms the SVM approach, as expected
- Similar patterns of model performance is observed when comparing across different test sets.
- Better scores on Wikipedia test set than OLID
- Poor performance on tweets about different subject matter.

SVM

Test Data Name	Precision	Recall	F1-score
OLID Test	0.806	0.718	0.743
Reddit	0.714	0.676	0.673
Wikipedia	0.864	0.863	0.862
Textgain	0.508	0.512	0.500

BERT

Test Data Name	Precision	Recall	F1-score
OLID Test	0.836	0.798	0.814
Reddit	0.719	0.705	0.707
Wikipedia	0.908	0.903	0.903
Textgain	0.546	0.590	0.500

Conclusion

- Further training of pre- trained language models on domain specific data results in significant improvements in model performance.
- Testing models in cross-domain as well as cross-topic settings provide useful insights into generalizability.
- *Is there “one” social media, Twitter, Facebook offensive language?*

References

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In Proceedings of the 5th Italian Conference on Computational Linguistics, pages 1–6.

Tommaso Caselli, Valerio Basile, Jelena Mitrovic, and Michael Granitzer. 2020. Hatebert: Retraining BERT for abusive language detection in english. CoRR, abs/2010.12472.

Ilia Markov and Walter Daelemans. 2021. Improving cross-domain hate speech detection by reducing the false positive rate. In Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 17– 22, Online, June. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. CoRR, abs/1902.09666.

Thank you for your attention.

For questions and remarks:
elif.kilik@student.uantwerpen.be