Offensive Language Detection

Bridget Tyree, Yi-Chien Lin,
David Yi, Levon Haroutunian

Outline

- Task description
- System architecture
- Core approach
- Issues and successes
- Related Readings

Task Description

Offensive Language Detection

- Primary Task:
  - Subtask A of OffensEval 2019
  - Binary classification: is an English tweet offensive (OFF) or not offensive (NOT)?
  - Data: OLID (Offensive Language Identification Dataset)
    - Total: 13,240 tweets – 8,840 NOT & 4,400 OFF tweets
- Adaptation Task:
  - Subtask A of OffenEval 2020
  - Detect offensive tweets in Arabic, Danish, Greek, and Turkish.

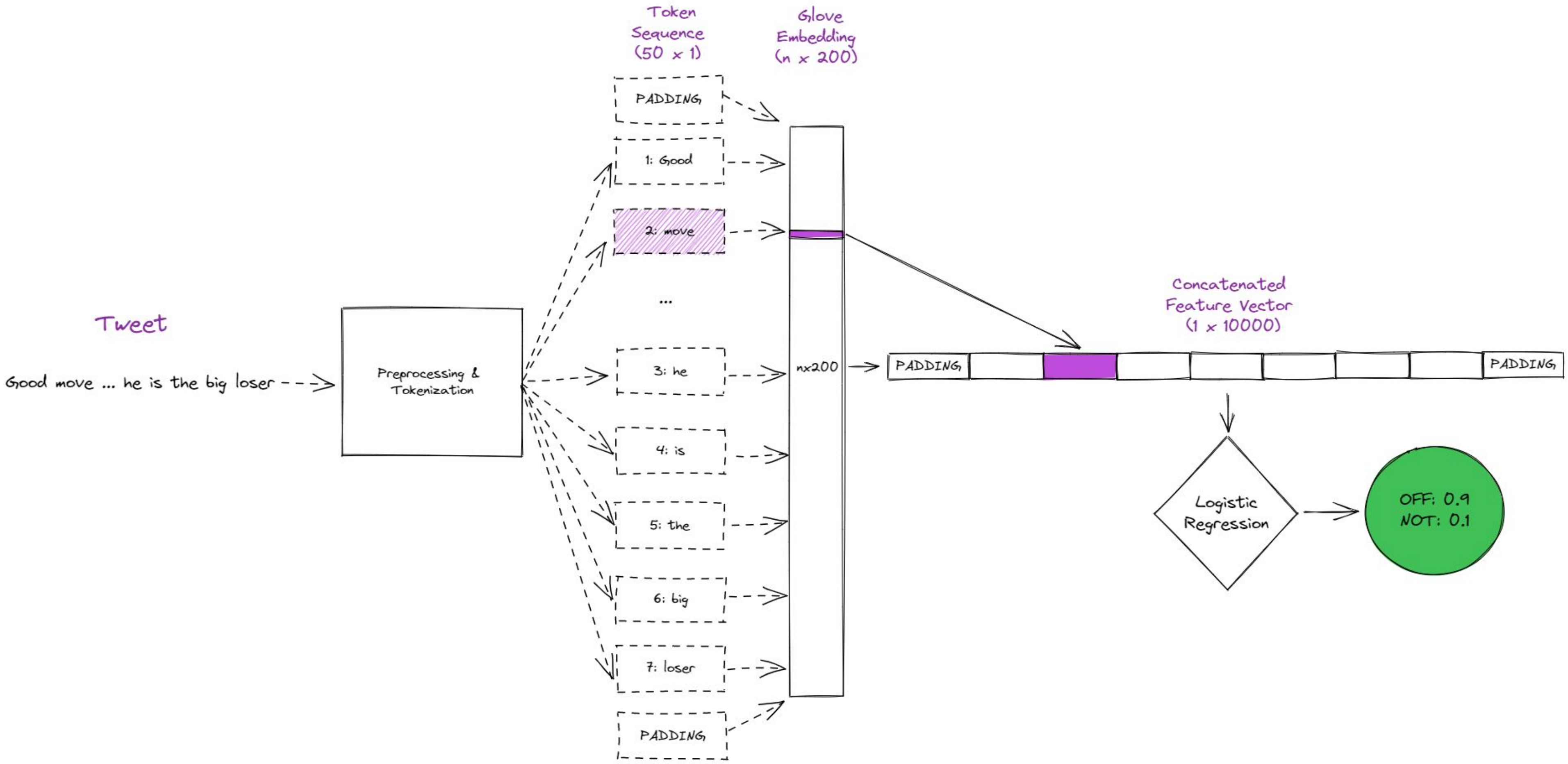
Tweet	**A**	**B**	**C**
@USER He is so generous with his offers.	NOT	—	—
IM FREEEEE!!!! WORST EXPERIENCE OF MY FUCKING LIFE	OFF	UNT	—
@USER Fuk this fat cock sucker	OFF	TIN	IND
@USER Figures! What is wrong with these idiots? Thank God for @USER	OFF	TIN	GRP

Table 1: Four tweets from the OLID dataset, with their labels for each level of the annotation model.

From Zampieri et al., 2019

System Architecture

System Architecture



Core Approach

Featurizing the Tweets

Tokenization: Keras tokenizer that converts tokens to unique integers with minimal preprocessing

- Pad and truncate all sequences to 50 tokens, since our classifier only takes in fixed-length inputs

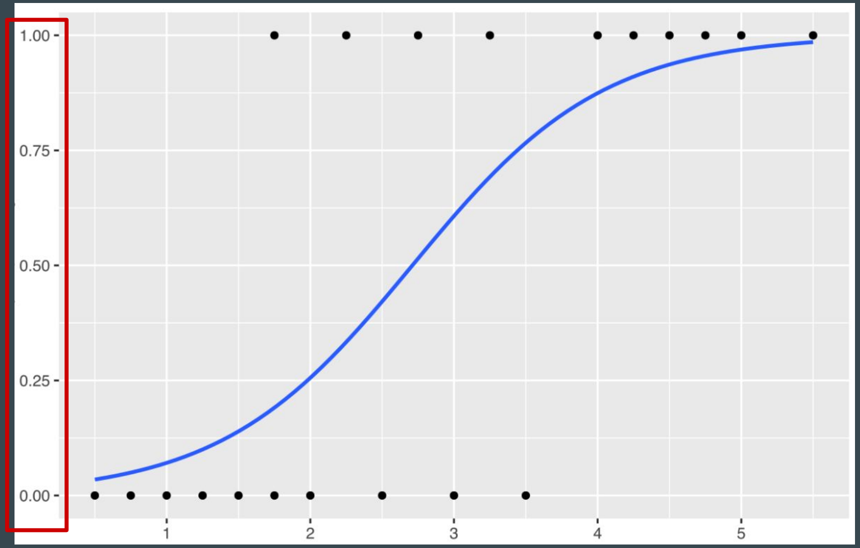
Word Embeddings: 200-dimensional GloVe Embeddings trained on Twitter Corpus

Featurization: Get GloVe embedding for each token in the sequence and stack them horizontally

- Input: Sequence of length 50
- Output: 10000 (50 * 200) dimensional vector

Binary Classification

- Logistic regression
  - Uses logistic function to model a binary output
- Scikit-learn parameters
  - L2 regularization
  - Class weight: balanced
  - Optimization problem: limited-memory BFGS



From Wikipedia

Issues and Successes

Results

Macro F1-Score:

0.5980265654648956

	F1 Score
SVM	0.690
OUR MODEL	0.598
ALL NOT	0.420
ALL OFF	0.220

Confusion Matrix		Predicted label	
		negative	positive
True label	negative	667	210
	positive	254	193

Error Analysis

- Untargeted and targeted offensive tweets were equally likely to be incorrectly classified. This suggests that our model is not making use of word identity.
- Upon looking at the data further, it seems like “controversial” words (like MAGA, antifa, gun control) are not strong indicators.
    - May also be true of negative sentiment more broadly.
- Our model may also be missing helpful information about sequence, which can be useful for determining whether a statement is a threat or insult.

Potential Next Steps (David)

Current System

- Classification: RNNs, (bi)LSTMs
- Emoji Embeddings (emoji2vec)
- Rule-based expansion of acronyms and abbreviations
- Incorporate syntactic or semantic parse trees during preprocessing/tokenization

Major Changes

- Used fine-tuned pretrained Language Model (i.e. RoBERTa) to create contextual word embeddings (Barbieri et al., 2020)
- Classification Ensemble

State of the Art (2019)

Method	F1 Score
BERT (fine-tuned)	0.829
CNN	0.800
BiLSTM	0.750
SVM	0.690
GloVe + Logistic Regression	0.598

SemEval-2019 Task 6: (Zampieri et al., 2019)

Related Reading

References

- SemEval 2019 Report: Zampieri et al., 2019
- TweetEval: Barbieri et al., 2020
- Abusive language overview: Talat et al., 2017
- GloVe Embeddings: Pennington et al., 2014