Hate Speech Detector

Yuxiang Wang, Jingxi Liu, Wenkai Luo, Yuetong Liu

Application: Text Data

Problem Definition

We are trying to...

1. Decide whether a post on social media is a hate speech

2. Distinguish hate speech and other offensive language

Data we used: Text Data-more than 20,000 English Sentence with labels.

Supervised Learning

Classification

Methods for feature extraction:

LIWC dictionary, skip-grams, LDA topics, n-gram

Methods for classification models:

Logistics Regression and SVM, Convolutional Neural Networks

Why is this an interesting problem?

Why unique?

No legal definition of “hate speech” in US law

Statistics do not represent context.

Why important?

Racial inequality, Climate of intolerance….(ethical implication)

Motivation:

Internet makes the detection and supervision of hate speech possible to achieve by artificial intelligence.

Our Task:

Build a classifier that helps the system to detect hate speech on the internet.

Dataset(s)

Example:

count	hate_speech	offensive_la	neither	class	tweet
3	1	2	0	1	" bitch get up off me "
3	0	3	0	1	" bitch nigga miss me with it "
3	0	3	0	1	" bitch plz whatever "

24,783 data entries, 6 features, label with 3 classes

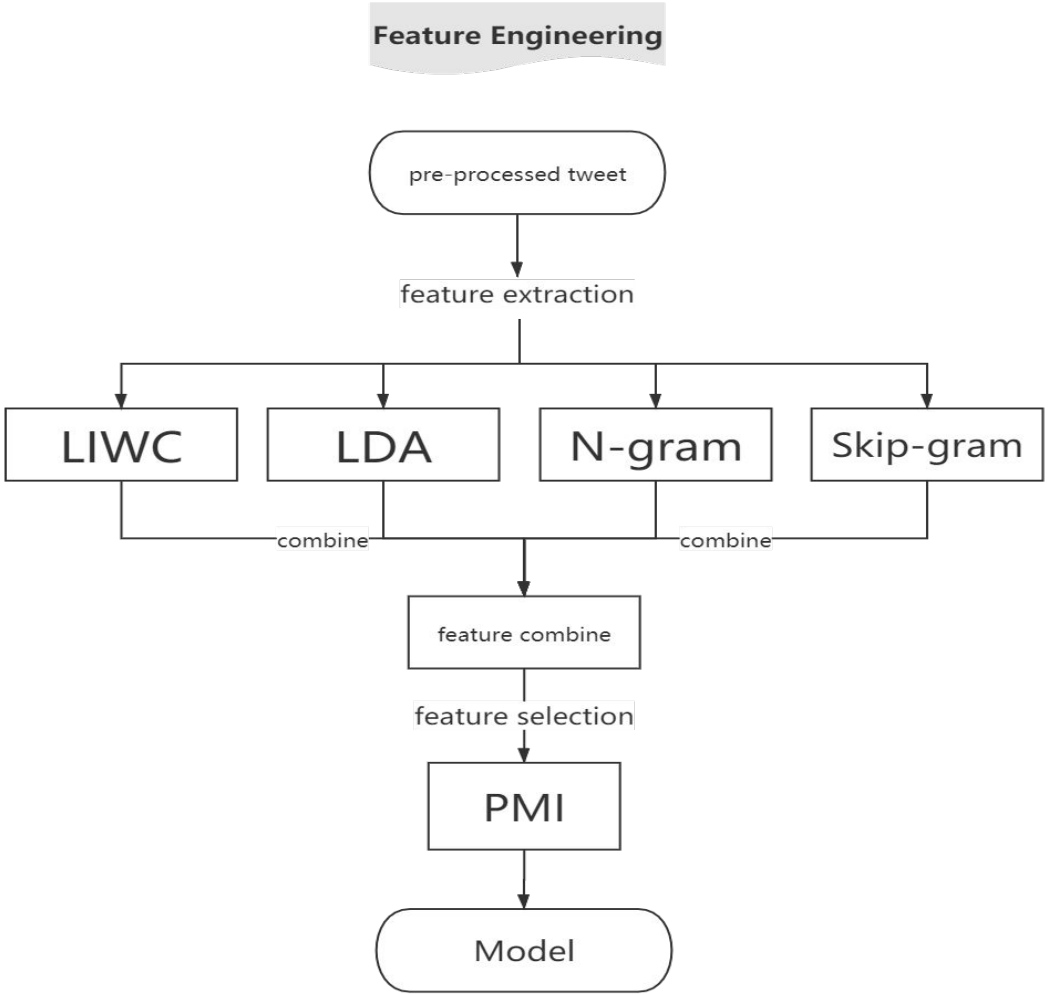
Data Pre-processing:

Delete duplicates, stop words, punctuation, and excessive whitespace

Convert tweet content to lowercase

Transform tweet(string) into a digestible form(list of words)

Features


Feature Engineering
pre-processed tweet
feature extraction
LIWC
LDA
N-gram
Skip-gram
combine
combine
feature combine
feature selection
PMI
Model

Methods

Baseline: SVM model with Several features combination

Method:

- SVM with new feature set (Skip-gram, LIWC, LDA, ngram)
- Feature Selection : PMI and Logistic regression estimator
- TextCNN

Training method: 5-fold cross-validation experiment

Previous Work: Classic Method vs Deep neural network

Results

Model	Precision	Recall	F1
SVM_skip	83.6	83.6	83.6
SVM_skip+fs	82.8	82.8	82.8
SVM_skip+lda	**85.4**	**85.4**	**85.4**

Models	Precision	Recall	F1
SVM (also [7])	86.6	86.4	86.5
SVM_{fs}	89.5	89.4	89.4
SVM+	86.2	86.4	86.3
$SVM+_{fs}$	89.5	89.7	89.6
$CNN+LSTM_{base}$, emb-learn	93.3	93.3	93.3
$CNN+LSTM_{base}$, emb-ggl1	93.3	93.3	93.3
$CNN+LSTM_{base}$, emb-ggl2	92.7	92.4	92.6
CNN+LSTM, emb-learn	93.4	92.9	93.1
CNN+LSTM, emb-ggl1	**94.2**	93.9	**94.1**
CNN+LSTM, emb-ggl2	94.0	**94.1**	94.0

Table 7: Results against baselines on the DT dataset

	precision	recall	f1-score
0	0.30	0.14	0.19
1	0.94	0.89	0.91
2	0.57	0.86	0.69

Deliverables

Deliverables:

Must: Data Preprocessing, Feature Extraction, Feature Selection

Expect: Prediction, Model Selection (on going)

Difficulties: Improving F1 value

Changes: PMI to L1 regularization

Modifications: Might not have enough time to work on topic clustering.

What we've learned

Relevant Concepts: SVM, DL, Classification

Surprising point: Data is imbalanced

Improvement: Collection information about user

Questions: How to reduce racial bias

Feedback: Applicable to real life

References

[1] Ward, K. (1998). Free Speech and the Development of Liberal Virtues: An Examination of the Controversies Involving Flag-Burning and HateSpeech. Retrieved from https://repository.law.miami.edu/umlr/vol52/iss3/4/

[2] Muntarbhorn, V. (2011). “Study on the prohibition of incitement to national, racial or religious hatred: Lessons from the Asia Pacific Region.”Retrieved from https://www.ohchr.org/Documents/Issues/Expression/ICCPR/Bangkok/StudyBangkoken.pdf

[3] Gershgorn, D. (2018).“Mark Zuckerberg just gave a timeline for AI to take over detecting internet hate speech.” Retrieved fromhttps://qz.com/1249273/facebook-ceo-mark-zuckerberg-says-ai-will-detect-hate-speech-in-5-10-years/

[4] Davidson, T. & Warmsley, D. & Macy, M& Weber, I. (2017) “Automated Hate Speech Detection and the Problem of Offensive Language”. Retrievedfrom https://arxiv.org/pdf/1703.04009.pdf

[5] Davidson, T. (2017) Data.world: Hate Speech and Offensive Language. Retrieved from https://data.world/thomasrdavidson/hate-speech-and-offensive-language

[6] Bower, J.M. & Beeman, D. (1995)The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.NewYork: TELOS/Springer–Verlag.

[7] Package: pandas, numpy, sklearn, seaborn, matplotlib, nltk, string, re, gensim, warnings, liwc