USC University of Southern California

Comparative Analysis of Detecting Offensive Language on Social Media

Viterbi School of Engineering,USC

Rehan Ahmed, Prerak Tusharkumar Pradhan , Mauli Amrishbhai Trivedi , Jimi Cao, and Romina Mir

CSCI554 - Fall 2021
Mohammad Rostami
Xuezhe Ma

USC

Offensive Language in Social Media

With the increase of social media usage, the focus on improving the social space for the community members has been expanded. The proliferation of social media enables people to express their opinions widely online. However, at the same time, this has resulted in the emergence of conflict and hate, making online environments uninviting for users. Nowadays, many people on the internet publish content containing offensive language on social media such as Facebook, Twitter, etc. In online comments, offensive and offensive language can lead to a host of different problems, including cyber-bullying that targets individuals (celebrity, politician, and product) and a group of people (specific country, age, and religion).

Offensive Language

- Offensive language - anything that causes offense, arousing a visceral reaction of disgust, anger, or hatred.
- Detect and analyze offensive language automatically.
- Can not be resolved with word matching

What We do?

Most existing offensive language detection techniques that are in place rely on manual human intervention and are particularly ill-suited for large datasets. In this project, we will conduct a study on various learning models on Offensive and Abusive Speech on Social Media Platforms and discuss the possibility of using additional features and context data for improvements. We will try to account for heterogeneity in this dataset by separately annotating both the comment as a whole and the individual sentences that comprise each comment and evaluate the best system for abuse and offensive language detection for large-scale datasets. We will also work on evaluating the performance of different abuse detection models in different languages.

Natural Languages

Natural Language Datasets:

English	24,783 tweets
Filipino	24,232 tweets
Chinese	8,969 comments
Korean	8,367 comments

Machine Learning techniques.

Traditional Algorithms

1. Naive Bayes
2. Logistic Regression
3. SVM
4. Random Forests
5. Gradient Boosting

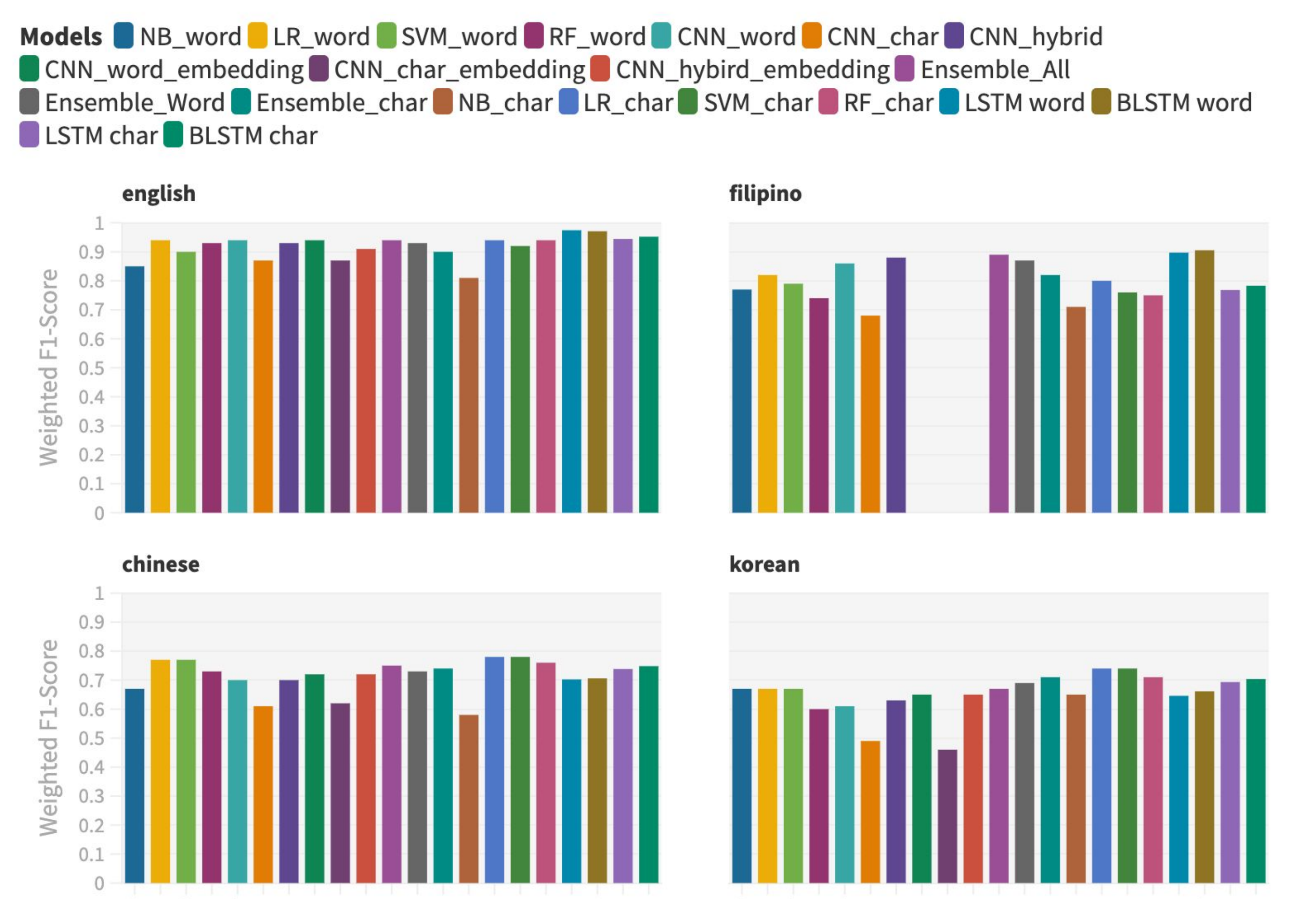
Neural Networks Algorithms

1. CNN
2. RNN/BiLSTM

Embedding:

- English
  - Glove
- Filipino
  - [None]
- Chinese
  - Pre-Trained Word Vectors (FastText)
- Korean -
  - Pre-Trained Word Vectors (FastText)

Results



Models
NB_word
LR_word
SVM_word
RF_word
CNN_word
CNN_char
CNN_hybrid
CNN_word_embedding
CNN_char_embedding
CNN_hybird_embedding
Ensemble_All
Ensemble_Word
Ensemble_char
NB_char
LR_char
SVM_char
RF_char
LSTM word
BLSTM word
LSTM char
BLSTM char
english
filipino
chinese
korean
Weighted F1-Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Main Challenges:

- Subjectivity and context-dependent nature.
  - Misclassification
- Ambiguity in the datasets
  - Emojis/slang

- Current Models:
  - Need human intervention
  - Major research focused on English
  - Heavy use meta-data

Conclusion:

- Need multilingual offensive language detection
- Current methods are not scalable
- Compared 5 traditional methods and 2 deep learning methods on 4 languages
- Neural network word-based models worked best for English and Filipino and traditional character-based models worked best for Chinese and Korean

Resources:

- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.
- Younghun Lee, Seunghyun Yoon, Kyomin Jung. 2013. Comparative Studies of Detecting Abusive Language on Twitter
- Wenliang Dai, Tiezheng Yu, Zihan Liu, Pascale Fung. 2020. Kungfupanda at SemEval-2020 Task 12: BERT-Based Multi-Task Learning for Offensive Language Detection
- Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, Yi Chang. 2016. Abusive Language Detection in Online User Content
- Ji Ho Park, Jamin Shin, Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection
- Michael Wiegand, Josef Ruppenhofer, Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets
- Georgios K. Pitsilis, Heri Ramampiaro, and HelgeLangseth. 2018. Detecting Offensive Language in Tweets Using Deep Learning. Thomas Davidson, Debasmita Bhattacharya, Ingmar
- Weber 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets.
- Davidson T, Warmsley D, Macy MW, Weber I. Automated Hate Speech Detection and the Problem of Offensive Language. CoRR. 2017
- Sepp Hochreiter, Jürgen Schmidhuber; Long Short-Term Memory. Neural Comput 1997.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1 (January 2014), 1929–1958.
- Zhang, Xianwei & Wu, Peng & Cai, Jiuming & Wang, Kun. (2019). A Contrastive Study of Chinese Text Segmentation Tools in Marketing Notification Texts. Journal of Physics: Conference Series. 1302. 022010. 10.1088/1742-6596/1302/2/022010.
- Eunjeong L. ParkO, Sungzoon Cho (2014). KoNLPy: Korean natural language processing in Python. Seoul National University, Industrial Engineering Department.
- Cho, Won Ik and Lee, Junbum, (2020). {K}orean Corpus of Online News Comments for Toxic Speech Detection - Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media.
- Neil Vicente Cabasag, Vicente Raphael Chan, Sean Christian Lim, Mark Edward Gonzales, and Charibeth Cheng (2019). Hate speech in Philippine election-related tweets: Automatic detection and classification using natural language processing.