Hate Speech Detection

Using Hurtlex Lexicon

Team 10

Miran M. Rashid Salim Fares

supervisor : Dr. Jelena Mitrovic

Universität Passau

July 23, 2020

Table of Contents

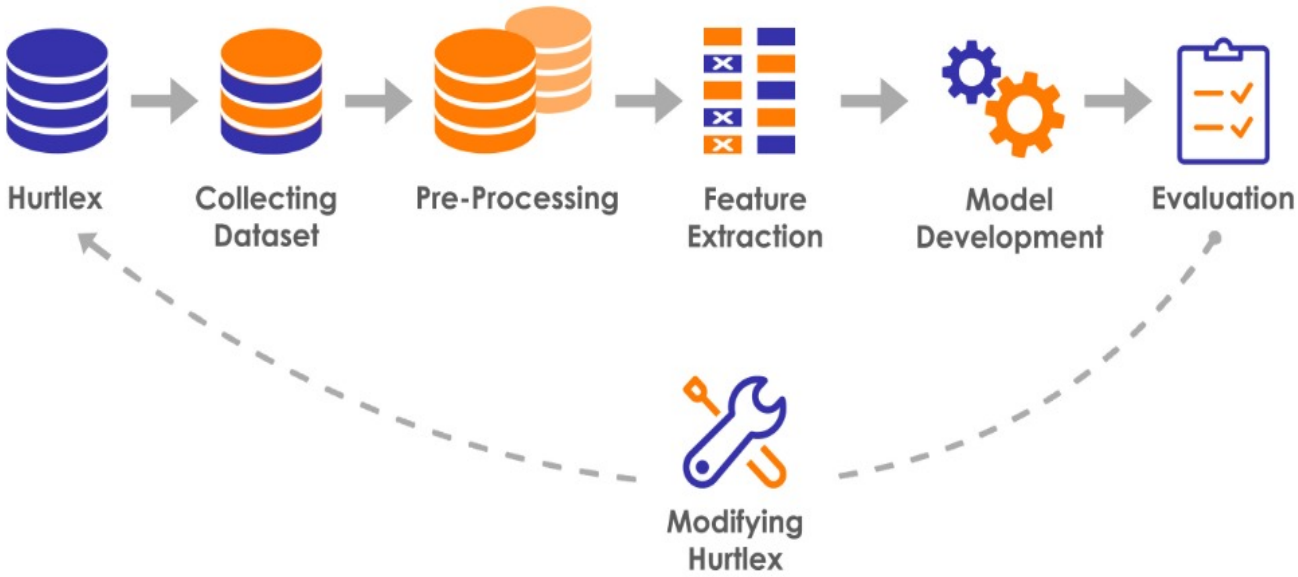
Table of Contents

Hate Speech

- Hate speech and offensive language
- Hurtlex lexicon
  - professions and occupations, physical disabilities and diversity, moral and behavioral defects, words related to social and economic disadvantage

Table of Contents

Methodology



Hurtlex Modification

- The Arabic lexicon consists of 3221 Arabic words.
  1. edited words: 120
     - الكائنات الطفيلية ← طفيلي
     - اللياقة
     - تبا
  2. deleted words:1181
  3. added words: 105

Hurtlex Modification

- The English lexicon consists of 8229 English words.
  1. added words: 65
     - refugee invasion, anti white, terrorist threat
  2. deleted words: 366
     - simple, love, academic, everyday

Dataset

AR-dataset contains tweets related to gender, sexual. orientation, religion, disability.

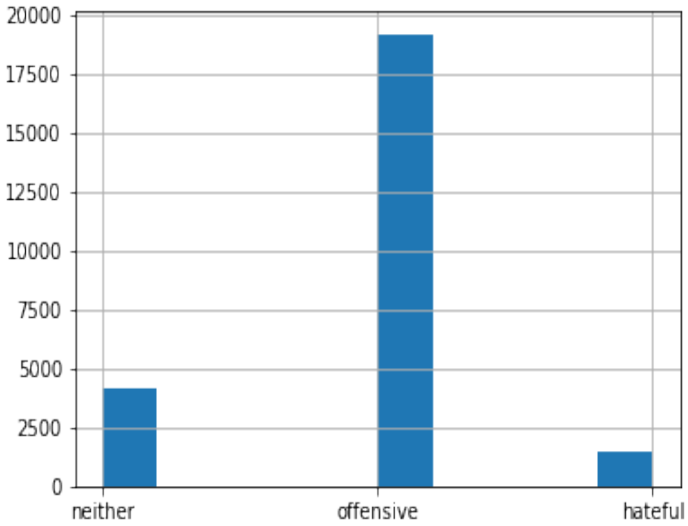


Figure: Ar-dataset

L-HSAB contain content that combines Syrian and Lebanese political tweets.

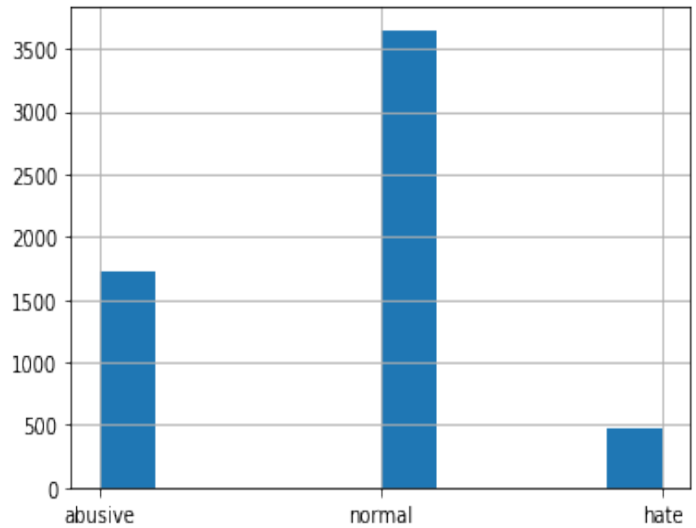


Figure: L-HSAB dataset

Dataset

contains news comments on Aljazeera.net which means the majority of it is also political.

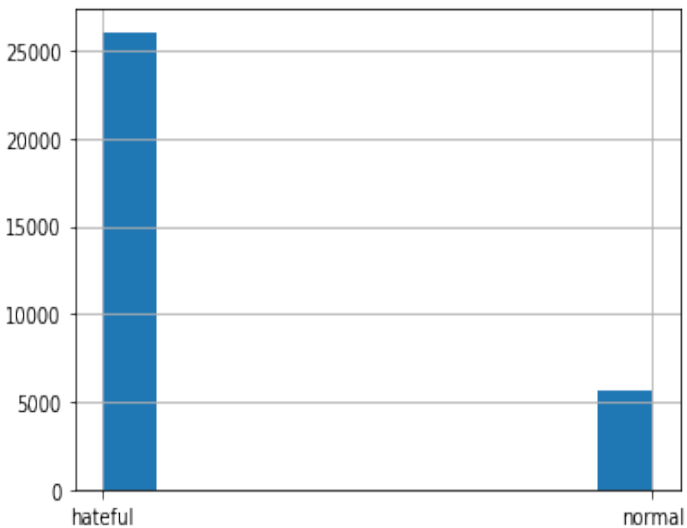


Figure: Aljazeera.net

En-dataset contain content that is racist, sexist, homophobic

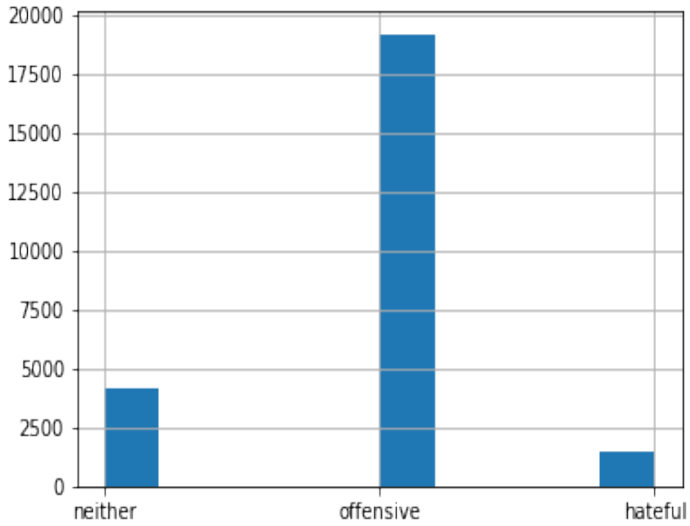
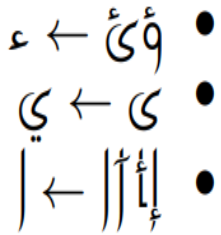


Figure: En-dataset

Pre-processing

- Data Cleaning:
  1. all non-Arabic characters(for the Arabic)
  2. numbers, symbols, punctuation, hashtags,emoji.
  3. Arabic stop words, English Stop words
- Normalization Arabic characters



Feature Extraction

- Count Vectors
- NLP based
- TF-IDF Vectors
- Word Embedding (only for Arabic)

NLP Based

- Word Count
- Character Count
- Average Word
- Punctuation Count
- Hate Word label
- Hate Words Count

TF-IDF Vectors

- $$TF(t) = \frac{\text{Number of times term t appears in a tweet}}{\text{Total number of terms in the tweet}}$$
- $$IDF(t) = log_e\left(\frac{\text{Total number of tweets}}{\text{Number of tweets with term t in it}}\right)$$
- $$TF - IDF \;\; = \;\; TF(t) * IDF(t).$$

$TF-IDF$ levels of input tokens (words, characters, n-grams)

Word Embedding (only for Arabic)

- Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation.

Models

- Naive Bayes Classifier
- Linear Classifier : Logistic Regression
- Support Vector Machine
- LSTM Just for Word Embeddings (only for Arabic).

Table of Contents

Evaluation Measures

TH (TN):correctly classified Hate (Normal) tweets.
FH (FN):incorrect classified Hate (Normal) tweets.

- Accuracy

$$\frac{TH + TN}{TH + TN + FH + FN}$$

- Precision

$$\frac{TH}{TH + FH}$$

- recall

$$\frac{TH}{TH + FN}$$

- F1 measure

$$2\,x\frac{PrecisionxRecall}{Precision + Recall}$$

Dataset Analysis Using HurtLex

	Before	After
Accuracy	0.6670	0.6676
Precision	0.6986	0.7188
Recall	0.8831	0.8262
F1-Score	0.7801	0.7688
Most Frequent Words	"اله", "اهل" , "قتل" “God”, “Parents”, “Killing”	"ارهاب", "بول", "قتل " “killing”, “Pi**”, “Terrorism”
Dominant Hate-Type	CDS, RE, AN	RE, CDS, AN

Evaluation The Word Embedding

- pre-trained AraVec model scored 0.69 Accuracy.
- our own Word Embeddings scored 0.71 Accuracy

Evaluation The Model

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
NB, Count Vectors	0.7417	0.6934	0.5288	0.5666
NB,WordLevel TF-IDF	0.7005	0.8206	0.3976	0.3853
NB,N-Gram Vectors	0.6931	0.8430	0.3862	0.3663
NB, CharLevel Vectors	0.6993	0.8026	0.4045	0.3994
NB,NLP Features	0.6577	0.3984	0.3595	0.3344
LR,Count Vectors	0.7314	0.6571	0.5916	0.6163
LR,WordLevel TF-IDF	0.7481	0.6882	0.5641	0.6010
LR,N-Gram Vectors	0.6995	0.6553	0.4382	0.4534
LR,CharLevel Vectors	0.7505	0.6862	0.5731	0.6085

Table: Model Evaluation For Arabic

Evaluation The Model

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
LR, NLP Features	0.6671	0.4825	0.3418	0.2871
SVM,Count Vectors	0.7152	0.6264	0.5917	0.6055
SVM,WordLevel TF-IDF	0.7386	0.6673	0.5733	0.6051
SVM,N-Gram Vectors	0.7002	0.6502	0.4533	0.4761
SVM,CharLevel Vectors	0.7455	0.6804	0.5770	0.6107
SVM,NLP Features	0.6664	0.4321	0.3379	0.2766
NB, Combined Features	0.9166	0.8033	0.8451	0.8206
LR,Combined Features	0.9875	0.9809	0.9674	0.9740
SVM,Combined Features	0.9837	0.9767	0.9614	0.9689

Table: Model Evaluation For Arabic

Evaluation The Model

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
NB, Count Vectors	0.8619	0.7721	0.5397	0.5682
NB,WordLevel TF-IDF	0.7946	0.5810	0.3757	0.3696
NB,N-Gram Vectors	0.7802	0.5679	0.3463	0.3171
NB, CharLevel Vectors	0.8070	0.8140	0.4090	0.4235
NB,NLP Features	0.7695	0.4274	0.3386	0.3063
LR,Count Vectors	0.8901	0.7333	0.6902	0.7042
LR,WordLevel TF-IDF	0.8913	0.7640	0.6661	0.6913
LR,N-Gram Vectors	0.7952	0.7228	0.3944	0.4050
LR,CharLevel Vectors	0.8921	0.7582	0.6648	0.6865

Table: Model Evaluation For English

Evaluation The Model

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
LR, NLP Features	0.7804	0.4556	0.3594	0.3435
SVM,Count Vectors	0.8841	0.7185	0.6890	0.7009
SVM,WordLevel TF-IDF	0.8907	0.7525	0.6775	0.6993
SVM,N-Gram Vectors	0.8023	0.6945	0.4238	0.4504
SVM,CharLevel Vectors	0.8940	0.7640	0.6706	0.6931
SVM,NLP Features	0.7758	0.4412	0.3429	0.3124
NB, Combined Features	0.8050	0.7872	0.7025	0.7290
LR,Combined Features	0.9875	0.9649	0.9529	0.9587
SVM,Combined Features	0.9872	0.9670	0.9512	0.9588

Table: Model Evaluation For English

Summary

- Hurtlex Modification
- Pre-processing and normalization
- Count Vectors, TF-IDF Vectors, NLP based, Word Embeddings.
- Naive Bayes Classifier, Linear Classifier : Logistic Regression, Support Vector Machine, LSTM
- LR,Combined Features