Structure-based hate speech detection

What is hate speech ?

- Hate speech is defined by the Cambridge Dictionary as "public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation".
- With the rise of social media and user generated content, detecting and classifying hate speech is becoming quite important. To automate the process of hate-speech detection,
- We look at a system that tries to utilize the grammatical structure of the system as features in order to classify a sentence as hate-speech or not, in order to avoid bias towards certain named entities

Datasets used

1. **Hate speech dataset from a white supremacist forum**
   - These files contain text extracted from Stormfront, a white supremacist forum. A random set of forums posts have been sampled from several subforums and split into sentences. Those sentences have been manually labelled as containing hate speech or not, according to certain annotation guidelines.

2. **Automated Hate Speech Detection and the Problem of Offensive Language**
   - Repository for Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language." ICWSM.

Preprocessing

- **Original sentence** : The only thing more disgusting than a White woman with a groid is a White woman who drags her White child into the filth with her.
- **Stopwords removal** : The only thing more disgusting woman groid White woman drags White child filth.
- **POS tagging** : TheDT thingNN disgustingVBG womanNN groidNN womanNN dragsVBZ childNN filthNN.
- **Stemming** : The onli thing more disgust woman groid White woman drag White child filth.

- **Abbreviations used** :
  - **N** - normal (without PoS tags, only stop words removed)
  - **P** - with PoS tags,
  - **P+S** - Stemming + PoS tagging

Phase I : Trying out models on dataset 1

- **Results**

Models used	Accuracy	Precision	Recall	F1 score
Naïve Bayes(P)	87.9	0.778	0.554	0.567
SVM (P)	87.2	0.973	0.503	0.473
Logistic Regression (N)	87.6	0.72	0.609	0.636
Decision tree (P+S)	85.4	0.65	0.609	0.624
CNN (single conv layer) (P+S)	85.6	0.447	0.199	0.276
CNN (complex model) (P+S)	82.5	0.368	0.372	0.370
LSTM (P+S)	82.8	0.337	0.355	0.293
BERT (N)	90.4	0.913	0.892	0.855

Note : BERT performed exceptionally well despite the skew.

Phase I: Experimenting with n-gram and LR models

- N-gram with Logistic regression (PoS tags only)

Models used	Accuracy	Precision	Recall	F1 score
Unigram/LR	86.2	0.701	0.587	0.835
Bigram/LR	86.5	0.733	0.555	0.825
3-gram/LR	86.4	0.733	0.541	0.818
4-gram/LR	86.3	0.732	0.536	0.816

- The N-Gram models show similar setbacks as well, with poor recall, but as far as precision and F1-score are concerned, they seem to perform better than their vanilla counterpart in the first run of experiments.

Phase I : Issues with previous models

- Although the accuracy seems to be very high, the precision, recall and the F1-score seemed to be terrible.
- Following are the reasons :
  - Highly skewed dataset.
  - 1437 labelled as hate speech and 9507 as non hate speech
- Confusion matrices

1908	0
279	2

SVM

1733	**153**
223	78

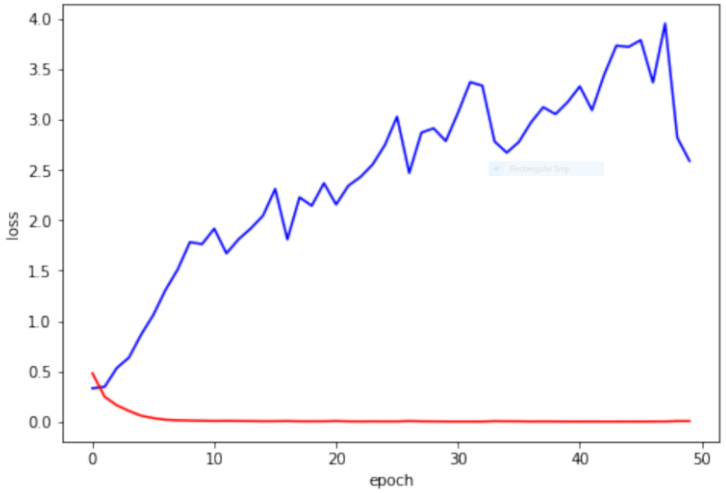
LSTM

Phase I : Issues with previous models

```
model = Sequential()
model.add(Embedding(vocab_size, 32, input_length=sen_len))
model.add(Conv1D(32, 3, padding='same', activation='relu'))
model.add(MaxPooling1D())
model.add(Conv1D(64, 2, padding='same', activation='relu'))
model.add(MaxPooling1D())
model.add(Flatten())
model.add(Dense(64, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(32, activation='relu'))
model.add(Dropout(0.2))
model.add(Dense(1, activation='sigmoid'))
model.compile(loss='binary_crossentropy',optimizer='adam', metrics=['accuracy'])

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.1, random_state=4)
```

The above CNN model was trained in batches of 50 points for 40 epochs. Although the model did show a few dips in validation loss, the trend is clearly upward (with the exception of the steep drop towards the end).



Phase II: Overcoming the imbalance problem

- **Naïve approach**
  - Oversampling : duplication of data points

Model used	Accuracy	Precision	Recall	F1 score
LSTM (N)	95.89	0.960	0.928	0.995
LSTM (P)	96.73	0.939	1.0	0.968
LSTM (P+S)	93.20	0.885	0.995	0.936

  - Undersampling : dropped out majority class samples randomly

Model used	Accuracy	Precision	Recall	F1 score
LSTM (N)	65.33	0.661	0.636	0.649
LSTM (P)	63.92	0.645	0.636	0.641
LSTM (P+S)	65.15	0.629	0.722	0.673

Phase II: Overcoming the imbalance problem

- **Synthetic Minority Oversampling Technique (SMOTE)** (unpreprocessed)

Model used	Accuracy	Precision	Recall	F1 score
SVM (P)	80	0.800	0.809	0.809
3-gram/LR (P)	82.9	0.833	0.828	0.828
CNN (complex)(P)	79.7	0.816	0.771	0.793
LSTM (U)	87.1	0.865	0.879	0.872
LSTM-attention (U)	87.7	0.897	0.851	0.873

- Conclusion : Clearly SMOTE has improved the scores of all the models, and hence it seems to have fixed the issues faced in the first experiment

Phase II: Experimenting with n-gram and LR models

N-gram with Logistic regression (PoS tags only) (Previous results)

Models used	Accuracy	Precision	Recall	F1 score
Unigram/LR	86.2	0.701	0.587	0.835
Bigram/LR	86.5	0.733	0.555	0.825
3-gram/LR	86.4	0.733	0.541	0.818
4-gram/LR	86.3	0.732	0.536	0.816

N-gram with Logistic regression (PoS tags only) (with SMOTE)

Models used	Accuracy	Precision	Recall	F1 score
Unigram/LR	82.3	0.830	0.822	0.821
Bigram/LR	82.8	.0833	0.827	0.827
3-gram/LR	82.9	0.833	0.828	0.828
4-gram/LR	82.6	0.830	0.826	0.826

Phase II: Trying out new dataset

- **About the dataset:**
  - The dataset has 19,190 sentences that are neutral, 1430 that are hate-speech and 4163 containing offensive language.
  - This dataset was obtained by making users tag a dataset on a crowdsourcing platform called CloudFlower. Each data point contains 5 columns
  - This dataset was picked up because it contains more number of data points than the Stormfront dataset, and the source this was picked up from contained annotations for dependency parsing of the sentence,
  - This would help us to experiment with Tree-LSTMs.
- **Preprocessing**
  - For the second dataset, before the preprocessing done in the first dataset, we additionally removed hashtags, '@' mentions and links to avoid any bias due to the same, since hashtags and mentions usually refer to or are related to some particular named entity.

Phase II: Trying out models on new dataset

- **Results without oversampling**

Models used	Accuracy	Precision	Recall	F1 score
Naïve Bayes(P)	86	0.751	0.518	0.548
SVM (P)	84	0.57	0.344	0.312
Logistic Regression (P)	90.4	0.735	0.687	0.699
Decision tree (P+S)	88.4	0.68	0.668	0.673
CNN (single conv layer) (P)	85.7	0.672	0.638	0.654
CNN (complex model) (P)	77.5	0.259	0.334	0.291
LSTM (P)	86.5	0.680	0.678	0.678
Bi-LSTM (P)	87.4	0.680	0.651	0.665

Phase II: Trying out models on new dataset

- **Results with oversampling (SMOTE)**

Models used	Accuracy	Precision	Recall	F1 score
SVM (P)	76.5	0.791	0.765	0.760
Logistic Regression (P)	86.1	0.865	0.860	0.861
CNN (complex model) (P)	66.4	0.669	0.673	0.671
LSTM (P)	68.5	0.686	0.685	0.685
Bi-LSTM (P)	68.3	0.681	0.682	0.681

Phase II: Experimenting with n-gram and LR models

N-gram with Logistic regression (PoS tags only)

Models used	Accuracy	Precision	Recall	F1 score
Unigram/LR	90.5	0.739	0.688	0.896
Bigram/LR	90.9	0.765	0.690	0.898
3-gram/LR	90.7	0.755	0.678	0.895
4-gram/LR	90.6	0.748	0.671	0.893

N-gram with Logistic regression (PoS tags only) (with SMOTE)

Models used	Accuracy	Precision	Recall	F1 score
Unigram/LR	83.4	0.854	0.834	0.835
Bigram/LR	78.2	0.827	0.781	0.774
3-gram/LR	78.0	0.824	0.779	0.771
4-gram/LR	77.7	0.823	0.776	0.767

Phase II: Experimenting with Tree LSTM and dependency parsing

- Tree LSTM incorporates non-linear semantic features such as Dependency Trees into our model.
- While using TreeLSTM, we can use the Dependency Parse Trees as features in our model. The difference between the standard LSTM unit and Tree-LSTM units is that gating vectors and memory cell updates are dependent on the states of possibly many child units.

- **Results**

Accuracy	Precision	Recall	F1 score
89.6	0.893	0.896	0.895

Error analysis

- Model : **CNN**
  - Misclassified sentences

Classified as not hate speech

Classified as hate speech

Sentences
httpabcnewsgocomInternationalprint id 2483106 AP Story
You nothing clownjust like Tricycle Sam
Actually They raised mother along father US government
Stop theorizing stop dreaming start practicing
The thing disgusting White woman groid White woman drags White child filth
Although evidence order kill Jews see see orders exterminate Bolsheviks Jews commissars etc
Definitely going
The concept honour respect feelings others
Just wondering hope take slogan way back Africa
Imagine steel cage battle royal match Wrestlemania multiply carnage 1000000 times get
Q If Timmy hop distance 2 feet makes 2 hops far hop starting point finishing point
I went chase bank make payment credit card chase banks city bullet proof glass front teller

Error analysis

- Model : **LSTM**
  - Misclassified sentences

Classified as not hate speech

Classified as hate speech

Sentences
If YOU teach child morals jew TV
One day true people come power italy go back came Italy Italians
I suppose concept male cow female cow producing MORE cows repeating process much understand
A apt analogy would went party showed find small group whites standing around darkies kept pouring door acting like owned place joined whites complaining
I hear jews looking open Holohoax center Antarctica next year
I remember classmate complaining hated jewish neighbors
Lookhow family growing two new fellas baby way
I Birmingham even nicer suburban area I getting darker darker
Today pay look table table guns priced retail
It important people realize gays want gay marriage desire sort government paperwork seeking right adopt defenseless children

Error analysis

- Model : **Bi-LSTM + attention**
  - Misclassified sentences

Classified as not hate speech

Classified as hate speech

Sentences
My friend lives Gdansk told almost everyone city white exception foreign students Chinese man owns dive bar
Deuteronomy 28 43 God speaking His children The stranger within thee shall get thee high thou shalt come low
Disgusting country become
Ready GO
CZ75 Tactical Sports 9 mm Mossberg shotgun 12 gauge Finnish RK 95 TP 762x39 M14 scope 762x51 H K 416 556x45 I sort regret able pick AKS74 carbine sixth
I clean toilet public toilets dirty work Economical contribution 2508
SEE UTV DID NOT POST MY COMMENTS TYPICAL SPREAD THE WORD THROUGHOUT IRELAND RUN THESE INVADERS OUT GO HOME AFGANIES WOULD YOU SEE US DOING THIS IN YOUR COUNTRY
Im going try hand laying
A lot white women trying create danger drawing people countries
YouTube Blacks Attack White Man On Subway YouTube White Girl Beaten On A Bus By Blacks YouTube Black gang attackTottenham High Road N1

Thank you