ML 8

Automated Hate Speech Detection On Social Media Data using Artificial Neural Networks

Muhammad M Rana
Faculty Advisor: Dr. Mehdi Allahyari
College of Engineering and Information Technology
GEORGIA SOUTHERN UNIVERSITY

Introduction

- Hateful contents are those that contain abusive speech targeting individuals (*cyber-bullying, a politician, a celebrity, a product*) or particular groups (*a country, LGBT, a religion, gender, an organization, etc.*).
- In 2016 , ***7,321*** hate crime related offenses were reported in the USA alone.
- In this work, we develop several machine learning based methods to detect hate speech on Twitter data.

Why This Research?

- Manual way of filtering out hateful tweets is not scalable and in fact impossible in real time.
- Social sites are facing the problem of identifying and censoring problematic posts, while weighing the right to freedom of speech at the same time. So, perfect prediction is a must
- Prediction is not straightforward. After observing the hate post ***'These bus drivers are all immigrant trash...'***. one may conclude any phrases with '***immigrant trash***' is a hate speech.But, ***'You should stop calling him a immigrant trash'.*** would definitely not considered as hate speech

Artificial Neural Network(ANN)

- For predicting the above complex scenario, word frequency based (Bag of Words) approaches are not sufficient. Our model needs to think like a human brain with memory.
- **ANN** is that model..
- ANN is inspired by the way how the human brain works. ANN models can learn through trial and error just like humans.
- In this experiment, prediction was attempted using two types of ANN. Convolutional Neural Network (**CNN**) and Recurrent Neural Network (**RNN**)
  - **CNN** learn to recognize smaller components first, then combine all the components to recognize large patterns.
  - Unlike CNN, **RNN** has a memory to track what was previously learned.

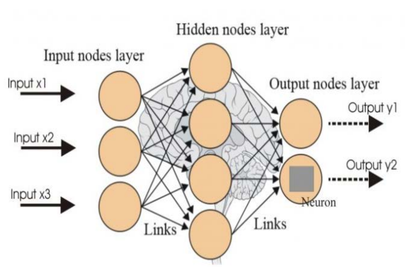
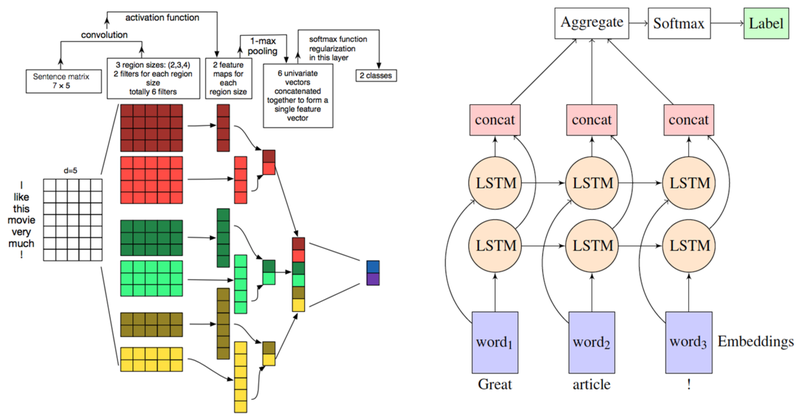


Fig: A Simple ANN model

Methodology

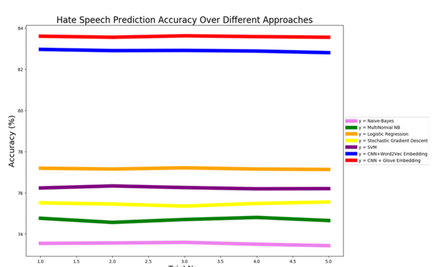
Figure: Hate Speech Detection Using CNN and LSTM(RNN)



- Level 1 : Data Fetching from Twitter using ***Tweepy.***
- Level 2 : Data Processing, cleaning, feature extraction.
- Level 3 : Experiment with baseline methods. We started with Bag of Words (BOW) models. Run the dataset on Naive-bayes, Logistic Regression, SVM, SGD models.
- Level 4 : Run the dataset on CNN model, where features are convoluted and pooled and later classified using softmax. Overfittings are prevented using dropout.
- Level 5 : Introduced ***Google's word2vec*** *and* ***Stanford's GloVe word embedding*** model for better semantic relation.
- Level 6 : Trying fit the data on ***LSTM,*** which can track previously recorded long sequences.
- Level 7 : Right now we are experimenting all the possible combinations of ANN for maximizing the accuracy.

Results

Prediction Model	Accuracy
Naive-Bayes	73.54
MultiNomial NB	74.76
Logistic Regression	77.19
Stochastic Gradient Descent	75.51
SVM	76.23
CNN+Word2Vec Embedding	**82.96**
CNN + Glove Embedding	**83.6**



- Experimental results show that, **CNN + Glove word Embedding** Technique gives the highest accuracy so far. It's **83.60% !!**

Sample Prediction

- We experimented with a dataset of 16K annotated tweets made available by the authors of [2]. Among these, 3383 are labeled as sexist, 1972 as racist, and the remaining are marked as neither sexist nor racist.
- Here is few sample predictions made by our code.

Sample Tweet	Hate Speech?	Correct Prediction
The girls should have less tickets on themselves and worry about the cooking. #MKR'	Yes	Yes
girls are pretty...awful. #gohome	Yes	Yes
Charlie Hebdo' editor killed in Paris terror attack	No	No
Islam is a religion with zero spiritual content.	Yes	Yes
The girl was raped by an Albanian	Yes	No

Our Ongoing Researches

- The LSTM(**RNN**) experiment is on progress. An accuracy of about **85%** is expected.
- Work is also being done on the character level CNN model.
- Attempts are also being made to introduce Hierarchical Attention Network (***HAN***), that can measure the importance of a word on a context.
- The best accuracy is expected to come from a ***hybrid of CNN and RNN model***.

Potential Uses

- The tool can be tuned by the use of a threshold which can be set by parents or teachers so ***online material can be filtered out*** before it appears on a web browser.
- A faster ***emergency response system*** can be made by classifying panic conversation on social media.
- A version of this application can be used for ***cyber crime prevention***.
- ***Online fraud detection*** is another significant use of this application.
- ***Hate crime prevention*** is also a noteworthy goal.

References

1. Convolutional Neural Networks for Sentence Classification by Yoon Kim
2. Deep Learning for Hate Speech Detection in Tweets by Pinkesh Badjatiya
3. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter by Zeerak Waseem et. al.