Graduate Capstone Project

STAT 5190

PROJECT PROPOSAL

Proposed title:

Identifying hate speech from tweets using deep learning

COURSE INSTRUCTOR:

DR NFI. NDIKINTUM

**INTRODUCTION**

**Hate Speech**

A hate speech is any speech made to arouse and propagate hate or prejudice against an Individual or a group. It is characterized by the use of insulting and derogatory words to stigmatize an the individual or group.

The following is considered hate speech:

* Any words that are uttered with the clear intent of inflicting injury or inciting and immediate break of the peace will be considered hate speech.
* Verbal utterances that are not words but inflict injury by their intent and known connotation, such as catcalls, will be considered hate speech.
* Likewise, printed paraphernalia such as posters and fliers which support racist beliefs or contain injurious statements.

Research in hate speech detection has increased over the years. In the paper “Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach” Aditya Gaydhani et al used logistic regression, support vector machine and Naïve Bayes’ algorithms in the detection of hate speech and offensive language on twitter.

In this research, I would be exploring deep learning algorithms in detecting hate speech on a collection of Nigerian Presidential Election twitter tweets. The collection process and scope of the data set is discussed below

**Deep Learning**

Deep learning is a class of [machine learning](https://en.wikipedia.org/wiki/Machine_learning) [algorithms](https://en.wikipedia.org/wiki/Algorithm) that use a cascade of multiple layers of [nonlinear processing](https://en.wikipedia.org/wiki/Nonlinear_filter) units for [feature extraction](https://en.wikipedia.org/wiki/Feature_extraction) and transformation. Each successive layer uses the output from the previous layer as input. (2)

Deep learning models are extensions of neural network models. In neural network, models consist of shallow nets that have – one input, one hidden layer and one output. Deep-learning networks are distinguished from these ordinary neural networks having more hidden layers, or so-called more depth. These kinds of nets are capable of discovering hidden structures within unlabeled and unstructured data (i.e. images, sound, and text), which constitutes the vast majority of data in the world.(3)

Deep learning algorithms surpasses natural language processing in handling complexity in language constructs and they outperform state-of-the-art char/word n-gram methods accorting to Pinkesh Badjatiya et al in “Deep Learning for Hate Speech Detection in Tweets”.

**Data Set and Data Collection**

Twitter has a feature that allows developers to download streams of tweets as they are being posted – and store them into a database. This feature is accessed through the twitter streaming API. Using twitter streaming API, I was able to gather about 1,000,000 tweets on Nigerian presidential election on twitter between January 5 – January 29, 2019. These are tweets about major presidential candidates in Nigeria. The data set is an unstructured textual and raw data with size of about 5gb. Since it is a raw data, there is need for it to pass through a data cleaning.

**Data Cleaning.**

Data extracted from Twitter usually come raw and unprocessed. For a data to be useful for analysis it needs to undergo the data cleaning. **Data preprocessing** is a **data** mining technique that involves transforming raw **data** into an understandable format. Real-world **data** is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data cleaning is an essential aspect of text analytics. A typical raw twitter tweet usually contains Emoticons, Irregular use of upper- and lower-case letters in a sentence, html website links, stop words, lots of punctuations. Etc. Some packages in python such has Lambda functions, Pandas and NumPy would be used in the data cleaning process.

**STATISTICAL METHODS AND MODELS**

Descriptive Statistics

After the data has been preprocessed and ready for analysis, some descriptive statistics would be performed. In text mining, analysis is being done by creating a numerical representation of the text – one way this can be done is tokenization (breaking sentences into word tokens). Examples of descriptive statistics that would be done include:

* Word frequency – frequency of which words occur
* Frequency of Frequencies: here we would check the frequency with which words appear and how many words occur that frequently
* Term Correlation: Here we would check for correlation between words
* Charts: different charts would be created for – most frequent words, popular words around each contestant, top location of tweeters, popularity of contestants, etc

**Modeling**

**Sentiment Analysis**

Sentiment Analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially to determine whether the writer's attitude towards a particular topic, product, etc., is positive, negative, or neutral.

In this proposed research work I plan to computationally categories every tweet around a presidential contestant into either of positive, negative or neutral sentiments. This can give us a feeling of the amount of hate speech geared towards a contestant.

Deep learning:

Multiple deep learning architectures including recurrent neural network and convoluted neural network would be employed in this project. Python programming language would be used as the main programming language. TensorFlow would be used as the framework to build the model.

**Objectives**

Our primary objective is to attempt to use highly efficient deep learning models in detecting or identifying hate speech from a collection of a large database of Nigerian presidential election tweets.

In addition, we hope to be able to answer the following research questions in the process of the research.

1. How are tweeters distributed by location?
2. Are some special tweeting agents disrupting the political environment with their tweets? (we can get this information by trying to find out who are the top tweeters.)
3. Who are the top contestants by popularity on twitter?
4. Words with highest correlation with each contestant and generating N-grams?

In text mining, an n-gram is a phrase or combination of words that may take on meaning that is different from, or greater than the meaning of each word individually.

We would generate n-grams using python to group words that have high correlation together.

1. Popularity of contestants by gender
2. What is the current political atmosphere?

Reference

1. <https://www.papermasters.com/hate-speech-us.html>
2. <https://en.wikipedia.org/wiki/Deep_learning>
3. <https://www.edx.org/course/deep-learning-with-tensorflow>