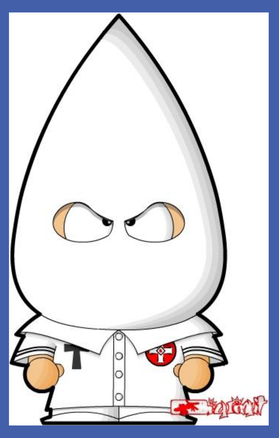


Racism Detection Algorithms

Hillel Merran, Elad Kapuza, Shir Shtinits

Why is Racism Detection Important?

Social media gets reports of thousands of posts each day. Going through them manually is a waste of precious time and resources. A smart, accurate classifier can automatically go over suspicious, possibly hateful posts and detect them without human intervention.

Aim of the Project

Comparing and developing three different methods for detecting racism in social media texts: SVM and Naïve-Bayes classifiers with a Bag of Words representation, and an MEMM-based Sentence-Document Model with feature representation.

Objectives

- Comparing between a smart learner and naïve ones
- Finding new ways to solve the racism detection problem
- Developing smart features to analyze the texts using a different take on sentiment analysis existing models

Data

Our data includes over 700 social media posts. The posts are divided into racist/not racist. All observations were collected and tagged manually.

20% of the posts were racist.

Challenges

- Posts are often written with sarcasm that a machine learner cannot detect, or are very subtle
- Posts contain slang, names or other words that are unfamiliar to the dictionary. Example: “Elor Azaria was right”
- Building an efficient Large-Margin classifier with a short runtime

Results

SVM: 79.7% accuracy

Naïve Bayes: 75% accuracy

Sentence-Document Model : 82.6% accuracy

Methodology

SVM:

- Simple Bag of Words
- Operating on entire posts without regarding the context within sentences – no individual sentence labels

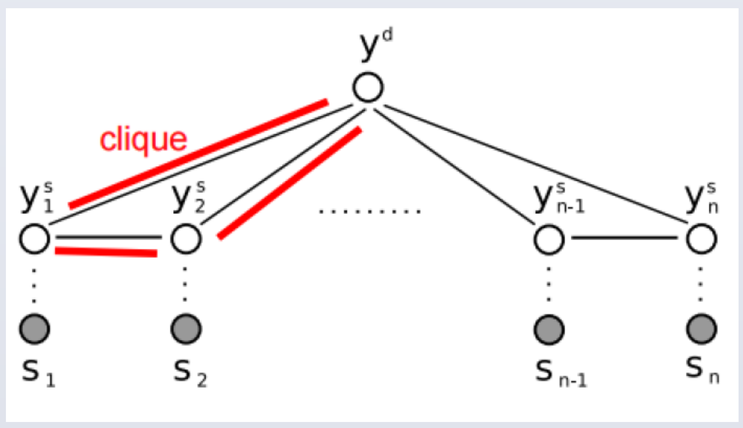
Naïve Bayes:

- Same as SVM
- Classifies using probabilities rather than vector distances

Sentence-Document Model :

- Using sentence-based cliques
- Determining clique scores and weights by features and the MIRA algorithm considering only k=4 highest scores output
- Determining posts’ and sentences tags using Viterbi algorithm
- The features used are: subjectivity, positivity/negativity, punctuation, racist/anti-racist dictionaries and more.

All models were tested using 5-fold Cross Validation.



The Sentence Document Model work method: each sentence’s label depends on its predecessor and on the post’s label

Conclusions

SVM tagged all posts as not racist. Possible explanations:

- It only had 142 racist posts to learn from, which might be too little for a naïve simple algorithm
- the weight of the racist posts and the penalty for mistakes we have chosen made it the optimal approach
- Bag of Words isn’t the right approach to tag something as short as a post

Therefore, SVM had a larger success rate than Naïve Bayes but it was by far the least useful classifier.

The Sentence Document had the highest success rate and performed successful learning.

- It had the most advanced learning systems
- It was the only algorithm that used sentences within their context
- Also the only one that used more criteria than the existence\absence of certain words

Future Use

The classifiers may be used by social networks to create features that highlight or hide suspicious posts.