Predictive Features for Hate Speech Detection using Image Classification of Hateful Symbols:

Final Project

Authors: Benjamin Cohen, [Haris](https://yu.instructure.com/courses/42439/users/19457) Sumra

Class: DAV6100 – Information Architecture

Professor: Brandon Chiazza

TA: Jacob Goodman

Date of Latest Modification: 7/27/2020

Version: 0.1

Contents

[Abstract: 2](#_Toc46763464)

[Introduction 2](#_Toc46763465)

[Characterizations of Hate Crime in Cyberspace 3](#_Toc46763466)

[Approaches taken to address the problems 4](#_Toc46763467)

[Other Dataset sources 5](#_Toc46763468)

[Solving the problem 7](#_Toc46763469)

[Resources and References 8](#_Toc46763470)

Abstract:

In recent years, hate speech has become a major issue in the domain of social media.

This paper introduces a method to detect hate speech in social media that contains hateful symbols. We first gathered hateful symbol data from different sources. This way, we created a hateful symbol dataset for this task. Then, we used this data for the training and evaluation of statistical models, which are based on state-of-the-art neural networks. Furthermore, we fine-tune pretrained descriptors that was used to define hateful symbols in our dataset. We also concluded our project by showcasing how these hateful symbols are offensive by adding expert knowledge to our trained model.

## Introduction

Hate crimes are nothing new to society; but what is hate speech? Hate speech is commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic (Nockleby, 2000). Our research is focused on the development of better methods for detecting hate crimes in the web domain and social media. In our project, we looked closely at hate symbols that are spread across web domain and social media. Over time there is more web content that is generated by users which also lead to an increase of hate speech. Suspects involved in hate-crime have an extensive social media history of hate-related posts, suggesting that social media contributes to their radicalization (Robertson, 2018). Multiple studies have looked at different ways to approach finding and systematic classifying hate speech. There are also multiple studies that have tackled related problems that can be helpful in for creating a new approach towards hate speech. So how is our approach different than other studies? We are providing users with the interface that will help user identify hate symbols across the web domain and social media. We are generating API and training models using neural networks. Goal of your work is to discover effective methods to improve upon hate symbols detections in the area of hate speech classification This paper provides a survey of the literature.

## Characterization of Hate Crime in Cyberspace

There are different ways that hate-crime is used negatively in the web.

1. **Extremist Groups**: Many extremist groups use the internet for a variety of reasons including communicating with members, organizing events, and educating others. But who are identified as Extremist Groups? An international Compilation of Terrorist Organizations, Violent Political Groups and Issue-Oriented Militant Movements. Even in the year 2000, according to the Simon Wiesenthal Center, there were approximately 2200 extremist websites. Schaer (2002) and Gertensfeld have done content analysis to identify the true nature of these websites. Gersteneld (2002) found that these websites are useful to recruit others and link related subgroups. For example, Kleg (1993) classified neo-Nazis, skinheads, Klan members, Identity Church members, and members of the Posse Comitatus all under the umbrella of ‘white supremacism.’ The internet allows these groups to form a collected identity according to Gertsenfeld. Being able to quickly identify extremist websites could potentially stop them spreading their ideas. More importantly, being able to tell when an extremist website would move words to violent actions could potentially save these events from occurring.
2. **Cyberbullying**:Cyberbullying is the defined as using electronic communication to bully a person. This is distinct from extremist groups in that it isn’t related specifically to any groups. Numerous studies have used NLP techniques to identify the sentiment in texts to identify cyber bulling (Kontosthathis, Edwards, & Leatherman).
3. **Hate Speech:** Hate Speech is using speech aggressively or threatening that expresses prejudice against a group. Nockleby (200) defines hate speech as “any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic.” Other social media platforms have similar definitions.

Approaching these types of issues requires a specific problem statement because the context in which a phrase is said can directly impact whether a phrase is considered negative speech. For example, a website that simply mentions the Ku Klux Klan is not hateful. At the same time, specifically pointing out a group to fit a stereotype would be. Warner and Hirschberg raise an example with the sentence “The next new item is a bumper sticker that reads: “Jew Bankers Get Bailouts, White Workers Get Jewed!” These are only 10 cents each and require a minimum of a $5.00 order”. The sentence itself is not hateful but it becomes hateful once Jews are mentioned.

Similarly, mentioning a word that disparages a group is not always hate speech. Consider this example also brought by Warner and Hirschberg: “Kike is a word often used when trying to offend a Jew.” This word has historical context and is not used to offend any group.

Some have chosen to limit the definition of anti-Semitic to focus on words that directly attacks certain groups. Facebook for example, defines hate speech as “Content that attacks people based on their actual or perceived race, ethnicity, national origin, religion, sex, gender or gender identity, sexual orientation, disability or disease is not allowed. We do, however, allow clear attempts at humor or satire that might otherwise be considered a possible threat or attack. This includes content that many people may find to be in bad taste (ex: jokes, stand-up comedy, popular song lyrics, etc.”( Banks, J 2010).

Fortuna and Nunes on the other hand define hate speech as “Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used.”(Fortuna and Nunes, 2018) This definition explicitly calls speech which uses humor as hate speech.

These different definitions are problematic because they lead to problematic datasets but as part of our research we will correctly identify hate symbols used on web domain and social media platforms and import those images to our dataset as will be discussed next.

## Approaches taken to address the problems

**Collecting the data**

There are 2 high level types of data that can be used to address this problem.

1. The first is to use computer vision techniques to analyze different videos.
2. Collect textual/image data.

We will focus on the second approach.

Because there is much contention about what hate speech is, it becomes harder to trust annotated data. Gertensfeld study was simply a content analysis and scraped data from websites. Davudsib created a dataset of 24,802 tweets by taking a hate speech lexicon from Hatebase and searching for tweets containing these terms. They then used crowdsourcing to annotate the data. If the annotation score was low, the data was excluded from the data set.

Twitter was not the only source used. Stormfront deGilbert took post from a white supremacist forum and hand annotated each post at the sentence level for whether it was hate speech or not. Warner and Hirschberg collected their data from yahoo forums before annotating it.

A third potential source of data used for similar problems is books. Basave, He, Liu, and Zhao (2013) created a model from Open Calais and Wikipedia documents.

## Other Dataset sources

**WaseemA** Waseem and Hovy also provide a dataset from Twitter, consisting of 16,914 tweets labeled as racist, sexist, or neither. They first created a corpus of about 136,000 tweets that contain slurs and terms related to religious, sexual, gender, and ethnic minorities. From this corpus, the authors themselves annotated (labeled) 16,914 tweets and had a gender studies major review the annotations.

· Waseem Z, Hovy D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In: SRW@HLT-NAACL; 2016.

**WaseemB** In a second paper, Waseem creates another dataset by sampling a new set of tweets from the 136,000 tweet corpus. In this collection, Waseem recruited feminists and anti-racism activists along with crowdsourcing for the annotation of the tweets. The labels therein are racist, sexist, neither or both.

· Waseem Z. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In: Proceedings of the first workshop on NLP and computational social science; 2016. p. 138–142.

**Stormfront** de Gilbert, et al. provide a dataset from posts from a white supremacist forum, Stormfront. They annotate the posts at sentence level resulting in 10,568 sentences labeled with Hate, NoHate, Relation, or Skip. Hate and NoHate labels indicate presence or lack thereof, respectively, of hate speech in each sentence. The label “Relation” indicates that the sentence is hate speech when it is combined with the sentences around it. Finally, the label “skip” is for sentences that are non-English or not containing information related to hate or non-hate speech. They also capture the amount of context (i.e., previous sentences) that an annotator used to classify the text.

· de Gibert O, Perez N, Garc’ia-Pablos A, Cuadros M. Hate Speech Dataset from a White Supremacy Forum. In: 2nd Workshop on Abusive Language Online @ EMNLP; 2018

**TRAC** The 2018 Workshop on Trolling, Aggression, and Cyberbullying (TRAC) hosted a shared task focused on detecting aggressive text in both English and Hindi. Aggressive text is often a component of hate speech. The dataset from this task is available to the public and contains 15,869 Facebook comments labeled as overtly aggressive, covertly aggressive, or non-aggressive. There is also a small Twitter dataset, consisting of 1,253 tweets, which has the same labels.

· Kumar R, Ojha AK, Malmasi S, Zampieri M. Benchmarking Aggression Identification in Social Media. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). ACL; 2018. p. 1–11

**HatEval** This dataset is from SemEval 2019 (Task 5) for competition on multilingual detection of hate targeting to women and immigrants in tweets. It consists of several sets of labels. The first indicates whether the tweet expresses hate towards women or immigrants, the second, whether the tweet is aggressive, and the third, whether the tweet is directed at an individual or an entire group. Note that targeting an individual is not necessarily considered hate speech by all definitions.

· CodaLab—Competition; Available from: <https://competitions.codalab.org/competitions/19935>.

## Solving the problem

After the studies got all the data, they used different approaches to build models. This became a supervised classification problem for the studies that used annotated data. Additional information from social media can help further understand the characteristics of the posts and potentially lead to a better identification approach. Information such as demographics of the posting user, location, timestamp, or even social engagement on the platform can all give further understanding of the post in different granularity.

The first step in this process of building the models was to create features. This section summarizes different NLP methods used to create features.

1. Dictionaries – Create a list of words and their frequency counts. The words can be classified into different types based on known connotations. For example, Shuhua Liu and Thomas Forss (2015) created content words for swears based on [www.noswearing.com](http://www.noswearing.com/). Other approaches including using an Ortony Lexicon to identify words with negative connotations that don’t necessarily contain profanity. This method is expanded with distance metrics to account for misspellings. For example, using edit distance on misspelled words can identify profane words that are purposely spelled wrong.
2. Bag of Words and N-Grams – With this approach, instead of having a predefined dictionary, this method creates a dictionary from the training data. This approach is combined with N-grams which allow to keep track of the context of each word. The disadvantage to this approach is that related words can be far apart in a sentence.
3. Frequency-Inverse document frequency is a measure of the importance of a word in a document within a document.
4. Template Based Strategy – This approach is to build a corpus of words and collect the K words occurring around each word.
5. Sentiment – Since hate speech has a negative sentiment, sentiments of phrases is often used as features.

After creating these features, they are fed into SVM, logistic regression, and Naïve Bayes models.

Aside from NLP, another approach used to tackle this problem is deep learning. Unlike NLP methods, deep learning methods do not rely on building a feature set. Deep learning creates the features through network structure. There are 2 main types of deep learning methods. The first is CNN and the second is RNN. CNN is useful for extracting word or character combinations while RNN is better for learning recurrent or orderly information. In one study, Zhang combined, both the CNN and GRU model (Recurrent Unit) to create a classifier.

## Resources and References

**Journals:**

<https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0072-6>- Us and them: identifying cyber hate on Twitter across multiple protected characteristics

<https://www.tandfonline.com/doi/full/10.1080/2330443X.2019.1660285-> Classifying Hate Speech Using a Two Layer Model

<https://ieeexplore.ieee.org/document/8963960-> Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets

Literature Review Based on these Journals:

2018

[A Survey on Automatic Detection of Hate Speech in Text](https://eclass.uop.gr/modules/document/file.php/DIT209/%CF%80%CF%81%CE%BF%CF%84%CE%B5%CE%B9%CE%BD%CF%8C%CE%BC%CE%B5%CE%BD%CE%B7%20%CE%B8%CE%B5%CE%BC%CE%B1%CF%84%CE%BF%CE%BB%CE%BF%CE%B3%CE%AF%CE%B1%202019-2020/A%20Survey%20on%20Automatic%20Detection%20of%20Hate%20Speech%20in%20Text.pdf)

2019

[DETECTION OF HATE SPEECH IN SOCIAL NETWORKS: A SURVEY ON MULTILINGUAL CORPUS](https://airccj.org/CSCP/vol9/csit90208.pdf)

Papers:

2020:

[Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification | Facebook AI Research](https://scontent-lga3-2.xx.fbcdn.net/v/t39.8562-6/78647855_818048661960903_4592063751666008064_n.pdf?_nc_cat=108&_nc_sid=ae5e01&_nc_ohc=hTRrI5ROWVYAX9hGCh7&_nc_ht=scontent-lga3-2.xx&oh=118a97abea1732f4e980857eb6183829&oe=5F2CE411) - ai.facebook.com

2019:

[Hate speech detection: Challenges and solutions](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221152)

[Automatic Identification of Hate Speech on Social Media Platforms using Machine Learning](http://trap.ncirl.ie/4172/)

[A Quantitative Approach to Understanding Online Antisemitism∗](https://arxiv.org/pdf/1809.01644.pdf)

2018:

[Hate Speech on Twitter: A Pragmatic Approach to Collect ...](http://ieeexplore.ieee.org/document/8292838/)

[Interpreting Neural Network Hate Speech Classifiers - ACL …](https://www.aclweb.org/anthology/W18-5111)

[Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter](https://arxiv.org/pdf/1803.03662.pdf)

2015

[Hate Speech Detection with Comment Embeddings](https://astro.temple.edu/~tuc17157/pdfs/djuric2015wwwB.pdf)

2012

[Detecting Hate Speech on the World Wide Web](https://www.aclweb.org/anthology/W12-2103.pdf)

2009

[Hate Online: A Content Analysis of Extremist Internet Sites](http://floodhelp.uno.edu/uploads/Content%20Analysis/Gertstenfeld.pdf)

Literature review resources

YU resources:

<https://library.yu.edu/literaturereview>

<https://www.yu.edu/sites/default/files/inline-files/Thesis%20Outline%202018-2019.pdf>

Texts referenced in above

<http://www.duluth.umn.edu/~hrallis/guides/researching/litreview.html>

<http://guides.library.vcu.edu/lit-review>

videos :

[https://www.youtube.com/watch?v=IClUgxoJf\_g](https://www.youtube.com/watch?v=IClUgxoJf_g&t=2s)

<https://www.youtube.com/watch?v=PzWTM4FApNg>

<https://www.youtube.com/watch?v=ES3TJrzaYKE>

From web

<https://www.scribbr.com/dissertation/literature-review/>

Resources:

<https://github.com/Hironsan/HateSonar>- HateSonar allows you to detect hate speech and offensive language in text, without the need for training. There's no need to train the model. You have only to fed text into HateSonar. It detects hate speech with the confidence score.

<https://www.perspectiveapi.com/#/home-> Perspective API uses machine learning models to score the perceived impact a comment might have on a conversation. Developers and publishers can use this score to give real-time feedback to commenters or help moderators do their job, or allow readers to more easily find relevant information.

<https://www.paralleldots.com/abusive-content-> Parallel dots identifies abusive and offensive language. It uses Long Short Term Memory (LSTM) algorithms to classify the text. It is trained on social media data and news data differently for handling casual and formal language.

<https://www.adl.org/resources/reports/the-online-hate-index#introduction-> The Online Hate Index (OHI) is a joint initiative of ADL’s Center for Technology and Society and UC Berkeley’s D-Lab, and is designed to transform human understanding of hate speech via machine learning into a scalable tool that can be deployed on internet content to discover the scope and spread of online hate speech.