Hate Speech Detection and High Performance Computing

Caroline Mansueti[1]; Dr. Randall Cone[1]
[1]Salisbury University Department of Mathematics and Computer Science

Abstract

Overt hate speech is relatively easy to detect compared to subtle hate speech detection, especially if it contains keywords that are current, well-known slurs since bag-of-word approaches can be applied. Subtle hate speech is trickier to flag since it can be masked with 'reclaimed' historical phrases, purposeful misspellings, and carefully crafted language. Creating an application that can successfully identify subtle and overt hate speech, as well as measure modern and historical hate speech levels present, is the focus of this project. We aim to recognize hate speech through a variety of natural language processing analyses.

Introduction

The World Wide Web has opened up many platforms of communication for users to share information, consume different opinions, and interact socially. Information and opinions can be disseminated via social media platforms such as Facebook and Twitter, community platforms such as Reddit and Voat, and news source websites and their comments section. In addition to these platforms, anyone can create a website with customized content. With respect to the former media mentioned, they have policies in place to monitor cyber-hate, but there is a lot of text to monitor and it is difficult to have an effective algorithm that captures all the subtleties that allow for hate speech to infiltrate dirty word filters. With respect to the latter type of websites, a personal website can be about any subject, regardless of whether that content is hateful or prejudiced. Being able to accurately identify online hate speech is critical. Hate speech can lead to psychological harm of groups and individuals, as well as bring about violent action against individuals belonging to a minority group.

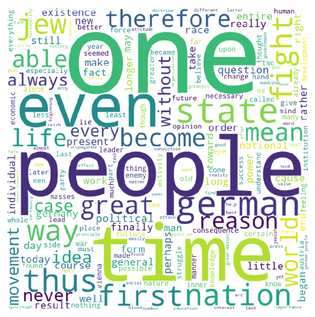
Results

A table is presented below outlining statistical reporting for textual data drawn from Daily Stomer articles, neutral Jewish articles, Mein Kampf Vol I, and Hitler's public speeches. The table is a result of the statistical data collected from processing the current contents of the three required corpora: common (neutral) speech, modern hate speech, and historical hate speech. Furthermore, some visualization is provided in the form of word clouds.

Hitler's Public Speeches

Neutral Jewish Content

Daily Stormer Content



Mein Kampf

Daily Stormer			Neutral Jewish			Mein Kampf			Hitler's Speeches		
Length	Freq	Rel. Freq.	Length	Freq	Rel. Freq.	Length	Freq	Rel. Freq.	Length	Freq	Rel. Freq.
3	10017	0.067941804	3	8264	0.058683179	3	3935	0.06382288	3	2585	0.064430098
4	25205	0.170956692	4	23697	0.168273873	4	9345	0.151569215	4	5760	0.143565713
5	23703	0.160769152	5	22731	0.161414247	5	10263	0.166458519	5	6887	0.171655741
6	24767	0.167985892	6	24018	0.170553314	6	10142	0.164495985	6	6900	0.171979761
7	20674	0.140224505	7	19556	0.138868374	7	8374	0.135820290	7	5627	0.140250741
8	15017	0.101855054	8	14826	0.105280349	8	6785	0.110047846	8	4087	0.101866852
9	11547	0.078319259	9	11118	0.078949610	9	5176	0.083951017	9	3231	0.080531392
10	7031	0.0476888120	10	6630	0.047080043	10	3442	0.055826778	10	2377	0.059245781
11	3579	0.0242751042	11	3627	0.025755553	11	1998	0.032406130	11	1104	0.027516761
12	1953	0.0132465154	12	2131	0.015132363	12	990	0.016057091	12	577	0.014381495
13	1098	0.0074473496	13	995	0.007065557	13	576	0.009342308	13	406	0.010119388
14	440	0.0029843659	14	459	0.003259387	14	225	0.003649339	14	122	0.003040801
15	128	0.0008681791	15	175	0.001242685	15	99	0.001605709	15	51	0.001271154
16+	106	0.0007189608	16+	74	0.000525478	16+	47	0.000762306	16+	37	0.000922210

Table 1: Lexical richness

Daily Stormer			Neutral Jewish			Mein Kampf			Hitler's Speeches		
Word	Freq	Rel. Freq.	Word	Freq	Rel. Freq.	Word	Freq	Rel. Freq.	Word	Freq	Rel. Freq.
jews	1826	0.0123851188	jewish	2004	0.0142305288	one	1070	0.0173546346	people	584	0.0145559681
jewish	1392	0.0094414487	said	1495	0.0106160881	people	493	0.007996107	german	518	0.0129109443
israel	1048	0.0071082171	jews	1307	0.0092810884	german	474	0.0076879409	germany	397	0.009895067
people	964	0.0065384745	mr	900	0.0063909560	would	432	0.007006731	one	356	0.0088731586
said	942	0.0063892562	anti	717	0.0050914616	even	389	0.0063093017	world	288	0.0071782856
jew	723	0.0049038559	one	698	0.0049565414	time	385	0.0062444246	must	262	0.0065302460
would	645	0.0043748092	people	625	0.004438163	state	295	0.0047846889	war	244	0.0060816031
one	644	0.0043680265	israel	622	0.0044168607	could	274	0.004440840	us	243	0.0060566785
anti	629	0.0042662683	would	583	0.0041399193	great	267	0.0043305490	state	222	0.0055532618
like	600	0.0040695899	like	537	0.00381327046	first	255	0.0041359176	would	215	0.00535878966

Table 2: Unigrams

Daily Stormer			Neutral Jewish			Mein Kampf			Hitler's Speeches		
Bigram	Freq	Rel. Freq.	Bigram	Freq	Rel. Freq.	Bigram	Freq	Rel. Freq.	Bigram	Freq	Rel. Freq.
anti semitism	197	0.00133	anti semitism	306	0.00217	great masses	63	0.00102	german people	136	0.00338
anti semitic	185	0.00125	anti semitic	195	0.00138	german people	57	0.00092	national socialist	39	0.00097
new york	140	0.00094	new york	179	0.00127	one could	50	0.00081	german nation	39	0.00097
united states	105	0.00071	united states	125	0.00088	german nation	46	0.00074	german reich	35	0.00087
donald trump	91	0.00061	jewish community	115	0.00081	looked upon	41	0.00066	one thing	34	0.00084
prime minister	83	0.00056	alt right	110	0.00078	self preservation	40	0.00064	world war	27	0.00067
white people	80	0.00054	year old	78	0.00055	point view	36	0.00058	economic life	27	0.00067
jewish community	77	0.00052	mr trump	77	0.00054	first time	36	0.00058	years ago	22	0.00054
pic twitter	66	0.00044	american jewish	68	0.00048	attitude towards	36	0.00058	rest world	21	0.00052
twitter com	65	0.00044	donald trump	67	0.00047	pan german	33	0.00053	fellow countrymen	21	0.00052

Table 3: Bigrams

Daily Stormer		Neutral Jewish		Mein Kampf		Hitler's Speeches	
Trigram	Freq	Trigram	Freq	Trigram	Freq	Trigram	Freq
pic twitter com	64	anti defamation league	39	pan german movement	27	one way another	7
new york times	45	new york times	33	instinct self preservation	24	whole german people	7
world war ii	43	world war ii	31	christian socialist party	10	twenty one years	7
prime minister benjamin	42	new york city	22	may looked upon	10	one thing certain	6
minister benjamin netanyahu	42	jewish community center	22	great masses people	10	national socialist germany	6
anti defamation league	40	prime minister benjamin	20	german workers party	10	german armed forces	6
israeli prime minister	28	minister benjamin netanyahu	20	let us say	6	right self determination	5
new york city	19	ku klux klan	14	new view life	6	national socialist movement	5
jewish daily forward	16	jewish community centers	12	half measures weakness	5	million square miles	5
president donald trump	14	anti semitic incidents	11	national instinct self	5	officers noncommissioned officers	5
god chosen people	14						

Table 4: Trigrams

Method

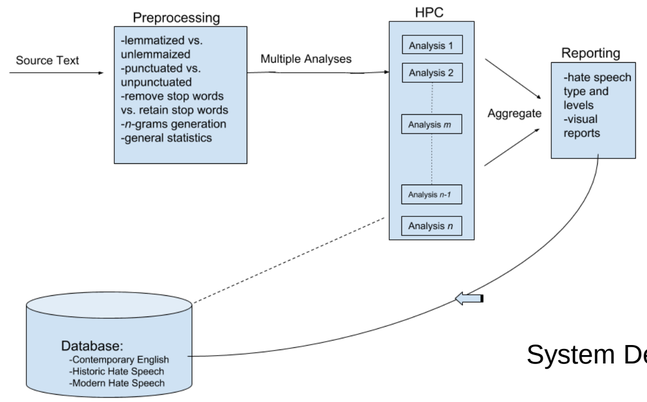
Custom software was developed using Python and a suite of libraries, most notably Natural Language Toolkit (NLTK). The capabilities of the software currently include: web scraping, text preprocessing, n-gram generation (for bigrams, trigrams, etc.), n-gram frequencies (general counts and relative frequencies), as well as visualization (word clouds and text dispersion plots).

Conclusions

Statistical analysis alone is inconclusive. Anti-Semitic articles from the Daily Stormer and Jewish articles that reported on similar subjects had overlapping vocabulary, but the difference of intent was not entirely captured.
The next step is to implement a Naïve Bayes classifier based on a balanced dataset. Such a dataset has been collected and is comprised of 500 anti-Semitic Daily Stormer articles and 500 articles from a variety of news sources that pertain to neutral Jewish content.

Direction of Future Work

We will continue to work on this project during the upcoming academic year. Pictured below is our system design, which requires more analysis modules to be created and combined. Potential analyses include: sentiment analysis for insecurity and/or culpability, proximal analysis, as well as building and testing classifiers with different baselines and features.



System Design