Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media

Joni Salminen[*†§], Hind Almerekhi[*], Milica Milenković[§], Soon-gyo Jung[*], Jisun An[*], Haewoon Kwak[*], Bernard J. Jansen[*]
[*]Qatar Computing Research Institute, Hamad Bin Khalifa University
[†]Turku School of Economics at the University of Turku
[§]Independent Researcher

Abstract

Online social media platforms generally attempt to mitigate hateful expressions, as these comments can be detrimental to the health of the community. However, automatically identifying hateful comments can be challenging. We manually label 5,143 hateful expressions posted to YouTube and Facebook videos among a dataset of 137,098 comments from an online news media. We then create a granular taxonomy of different types and targets of online hate and train machine learning models to automatically detect and classify the hateful comments in the full dataset. Our contribution is twofold: 1) creating a granular taxonomy for hateful online comments that includes both types and targets of hateful comments, and 2) experimenting with machine learning, including Logistic Regression, Decision Tree, Random Forest, Adaboost, and Linear SVM, to generate a multiclass, multilabel classification model that automatically detects and categorizes hateful comments in the context of online news media. We find that the best performing model is Linear SVM, with an average F1 score of 0.79 using TF-IDF features. We validate the model by testing its predictive ability, and, relatedly, provide insights on distinct types of hate speech taking place on social media.

Introduction

Hate speech, defined as hateful comments toward a specific group or target (Walker 1994), is rampant online. Several studies have reported the problem of toxic comments in social media (Djuric et al., 2015; Silva et al., 2016; Sood et al., 2012a). Hateful online comments have several drawbacks. First, they result in a vicious cycle of exchange of insults, known as 'online firestorms' (Pfeffer et al., 2014), making the comment threads toxic and counter-productive. Second, negative commenting has the potential to scare away high-quality discussants willing to contribute positively to the discussion (cf. Akerlof 1970), especially in online news media. Third, hateful comments spread hate and other negative emotions through emotional contagion (Kramer et al. 2014) and enhance the group polarization and echo chamber effects (Del Vicario et al., 2016).

While different methods have been applied to reduce hateful commenting, including counterspeech (Wright et al. 2017), non-anonymity, and mandatory registration (Hughey and Daniels 2013), these efforts have not been fully successful, and so online hate remains a highly topical research problem with major societal importance. One reason is that there is a lack of methods of understanding the types and targets of hate speech, without which it is difficult to determine how best to address the problem.

This open issue motivates our research to develop a granular taxonomy based on the open coding technique, where the classes emerge from the material (Glaser and Strauss 1967) and use it to train a model that learns to detect if a comment is hateful or not, and, if so, what group is being targeted. Automation is needed because manual moderation of thousands of comments is laborious and often neglected for that reason. In particular, media organizations producing dozens of videos per week on YouTube are facing real problems moderating hateful comments. Therefore, fully automatic or computer-aided moderation is needed to sustain the health of online communities. Toward this end of automatic detection of online hate, we present the following research questions:

1. How can hateful comments in social media be automatically detected and classified? *This question we answer by first developing a taxonomy of online hateful comments by qualitative open coding and then using data annotated using this taxonomy to train machine learning models.*
2. What are the common targets of online hate speech? *This question we will answer by applying the classifier to a dataset from a major online media company*.

Our aim with this research is part of the larger goal to automate classification and distinction of abusive and hateful language in social media. Several stakeholders are working on this problem, including Google's parent company Alphabet, with their Perspective API[1]. However, the accuracy of Perspective API and other solutions is not yet at a sufficient degree (Hosseini et al. 2017), and YouTube's comment sections, along with many other sites, remain usually with little moderation due to lack of automatic tools and the channel owners' resource constraints.

In particular, while the commonly used dictionary-based methods can be powerful indicators for hateful comments, alone they are not enough to detect all variants of hate speech (Saleem et al., 2017). Therefore, more granular models are needed, meaning, in practice, *multiclass classification*, where the hate is split into several subcategories according to its target and type of language used. In addition, hateful comments can contain overlapping targets and types of language (for example, at the same time being anti-Semitic and anti-government), prompting for *multilabel classification*. However, existing works using multilabel classification for online hate speech are extremely rare, and we could not locate prior work that had achieved good results. Therefore, aiming to fulfill that research gap is our goal, so that we create a multiclass, multilabel classifier that considers several categories of online hate. Through this effort, we are able to more accurately model the nature of hate taking place in online discussions. We aim to demonstrate not only how to train a machine to detect online hate, but also *what we, the researchers, can learn* from the comments annotated by the machine.

Related Literature

We queried academic databases, including Google Scholar and Science Direct, to identify related work. Search phrases included [+Online hate speech], ["Toxic comments"], and other topically relevant key phrases. We then manually evaluated the relevance to our research objective, finding several articles on offensive and hateful speech in social media, news sites, and other platforms for discussion and information sharing by users. Social media and news websites, such as Twitter, Facebook, Reddit, Yahoo! Buzz, Whisper, and YouTube have been the most common contexts for these types of analysis, especially Twitter.

Mondal et al. (2017, p. 87) define hate speech as "*An offensive post, motivated, in whole or in a part, by the writer's bias against an aspect of a group of people*." Davidson et al. (2017) distinguish between hate speech and offensive language. In some countries, this is crucial, since hate speech is a crime, and can result in imprisonment. In other countries, like the USA, where the freedom of speech is a constitutional right, removing hate speech represents a problem for social networks. In this work, we focus on the general concept of *hate*, not exclusively on *hate speech*.

Prior research identifies multiple challenges for automatic detection of online hate, summarized in Table 1.

Table 1: Challenges of automated detection of online hate speech.

Challenge	**Explanation**	**Reference**
Linguistic diversity	Language involves distractions, such as sarcasm and humor.	Saleem et al. (2017); Sood et al. (2012b)
Contextuality of hate	Hate speech can be contextually embedded, so that what in one community is perceived offensive is not so in another community.	Saleem et al. (2017)
Gaming the system	Users can subtly change their tone to fool the systems.	Hosseini et al. (2017)
Freedom of speech	Misclassification can result in limiting individuals' freedom of expression.	Mondal et al. (2013); Davidson et al. (2017)

Mondal et al. (2017) used a simple sentence structure "I <intensity> <userintent> <hatetarget>", allowing them to identify explicit hate targets. Intent is the emotion of the user; intensity is the level of emotion, and a hate target is the group receiving dislike or animosity. To avoid false positives, such as: "*I really hate owing people favors,*" they 1) placed a specific word before 'people' to specify hate targets (e.g., black people, Mexican people, stupid people), and, since not all hate contains the word 'people,' they 2) used 1,078 hate words from the Hatebase[2]. Using this strategy, they identified 20,305 Tweets, and 7,604 Whispers as hateful, most common categories on both social networks being race, behavior, and physical.

The study by Mondal et al. (2017) illustrates the limitations of using keywords only. The issue is the diversity of hate, which is not fully captured by the lexicon. Also, the method is susceptible for error, for example, it would find "*I hate police officers*", but miss "*police officers are dogs*." Saleem et al. (2017) further point out that a keyword used in one as a hate indicator may not represent hate in another community. For example, Sood et al. (2012b) used a profanity list with a stemmer, detecting 40.2% of profanity terms at 52.8% precision, concluding that even the best lists would not achieve reliable performance in profanity detection. Keywords are also prone to missing sarcasm and

[1] https://www.perspectiveapi.com/

[2] https://www.hatebase.org/, "World's largest and most authoritative structured repository of hate speech."

forms of humor, as these genres of language are particularly challenging to classify (Rajadesingan et al. 2015). Moreover, Nobata et al. (2016) note that the blacklist (a special collection of hateful words and insults) requires constant updates. Sood et al. (2012b) point out that adaptability to new terminology and slang is a major challenge, since the existing lists are missing the unfamiliar terms.

To overcome these concerns, Saleem et al. (2017) used labeled Latent Dirichlet Allocation (LLDA) to learn the topics, comparing to baseline language from Reddit, and ensuring that the chosen communities have distinct linguistic practices. This method showed a better performance than Naïve Bayes, showing that training a classifier on community-specific data may achieve a better performance than a generic keyword-based classifier.

Despite the shortcomings of dictionary-based methods, the use of language is crucial in detecting hate speech. To better model language, researchers have attempted applying word embeddings. Djuric et al. (2015) detect hate speech in comments collected from Yahoo! Finance, using 1) Paragraph2vec with Bag of Words (BOW) Neural Language Model, to discover masked insults and swearing; and 2) embeddings-based binary classifier to separate hateful and non-hateful comments. Paragraph2vec was able to discover some non-obvious swearing words and also obtaining better results than BOW models. In their context, most insults were targeting rich people (Djuric et al. 2015).

Previous studies have also found that using word embeddings (i.e., distributional semantics) performs well. For example, Nobata et al. (2016) detect hate speech, profanity, and derogatory language. They used N-grams, Linguistic, Syntactic, and Distributional Semantics, finding that combining all feature types gave the best performance for Finance and News contexts (Nobata et al. 2016).

Some of the more recent works utilize deep learning for hate speech detection. For example, Badjatiya et al. (2017) classified tweets using deep neural networks. Benchmark dataset of 16k tweets was analyzed, and 3,383 were labeled as sexist, 1,972 as racist, and remaining were labeled as neither. They found deep learning, e.g. convolutional neural network (CNN), better than the baseline methods (character n-grams, TF-IDF, BOW). The best accuracy was obtained when combining deep neural networks with gradient boosted decision trees (Badjatiya et al. 2017).

Park and Fung (2017) detected racist and sexist language through a two-step approach with convolutional neural networks. They used three CNN-based models, CharCNN, WordCNN, and HybridCNN, on Twitter data, containing 20k comments. The best performance was achieved with HybridCNN and the worst with CharCNN. When two logistic regressions were combined, they performed as well as one-step HybridCNN, and are better than one-step logistic regression (Park & Fung 2017).

Table 2 summarizes prior classifications of online hate.

Table 2: Hate Classifications from the Literature. N/A Indicates No Specific Target Was Mentioned in the Reference.

Source	**Category**	**Target**
Sood et al. (2012a, 2012b)	Politics, News	previous author, third party
	Business	
	Entertainment	
	Health, Lifestyle	
	World, Science	
	Travel, Sports	
Silva et al. (2016); Davidson et al. (2017)	Hate speech	
	Offensive	
Mondal et al. (2017)	Race	black and white people
	Behavior	insecure, slow, sensitive people
	Physical	obese, short, beautiful people
	Sexual orientation	gay people, straight people
	Class	ghetto people, rich people
	Gender	pregnant, sexist people
	Ethnicity	Chinese, Indian, Pakistanis
	Disability	retard, bipolar people
	Religion	religious people, Jewish people
	Other	drunk people, shallow people
Kwon and Gruzd (2017)	Public swearing	N/A
	Interpersonal swearing	N/A
Park and Fung (2017)	Sexism	N/A
	Racism	N/A
Badjativa et al. (2017)	Sexism	Female
	Racism	Pakistan
	Religion	Muslims
Saleem et al. (2017)	Hateful speech	Black
		Plus-sized
		Female

As seen from Table 2, notable exceptions of simple categories are Mondal et al. (2017) who employ 10 categories and Sood et al. (2012a) who classified hateful comments

under six categories, the target being either an author of a previous comment or a third party. Sood et al. (2012a) found political figures representing most targets of profanity, whereas lifestyle comments contained few insults. Their study demonstrates how automatic classification can provide more detailed information about hate in social media.

Overall, our literature review shows that 1) earlier taxonomies of hate targets tend to be coarse, and that 2) dictionary-based approaches alone are not sufficient in detecting and classifying hateful online comments. Granular classification is important e.g. to community managers and public policy makers who wish to understand online hate. To address these issues, we a) develop a granular taxonomy of online hate, and then b) use it to classify hateful online comments by their target and type.

Methodology

Research Context and Data Collection

We collect data from a major online news and media company with an international audience. This media company is highly active in social media, posting several videos per day on YouTube and Facebook and typically receiving thousands of comments per video. While exploring the social media presence of this online news media, we observed that many comments include hateful language, prompting us to find ways to detect and classify them using automated means. It seems logical that these comments, collected via YouTube and Facebook APIs, make a useful dataset for studying online hate speech and represent a real problem for online news publishers.

By using official APIs, we pull 137,098 comments from videos posted on YouTube and Facebook, in the period of July-October, 2017. The commentators are from 175 countries, although here we focus on English comments only. In the dataset, 79,439 (46%) comments are from Facebook, 57,659 (54%) from YouTube. YouTube Analytics does not provide country information at the comment level, but the aggregate numbers show that 38.3% of commentators are of **unknown origin** (a general limitation of YouTube data collection), 34.9% are from the **United States**, after which the relative share drops drastically, **India** being the second largest identified country with 4.2% of commentators. **United Kingdom** (3.6%) and **Canada** (3.0%) are the largest after India. It is likely that are most commentators are immigrants because the news organization is reporting on non-American issues and many commentators make references to their home countries, such as India (2,575 times mentioned), **Philippines** (642), **Pakistan** (1,231), etc. Therefore, the commentators are likely to be ethnically varied, from many countries in the world.

We explore the hate in the dataset by building a simple dictionary based on a) public sources of hateful words[3] and b) a qualitative analysis. We modified the hateful words from the public sources by looking at the data and identifying common terms associated with hateful comments. For example, 'hypocrites' was commonly used in hateful sense. We consider derogatory language in general (e.g., fucking), as well as specific targets (e.g., nigger, white devil, zionist) Overall, the dictionary includes 200 commonly appearing hateful words in this online news media[4].

Searching with that dictionary, we find that 22,514 comments (16.4%) contain these hateful wordings. Figure 1 illustrates the most commonly used hateful nouns.

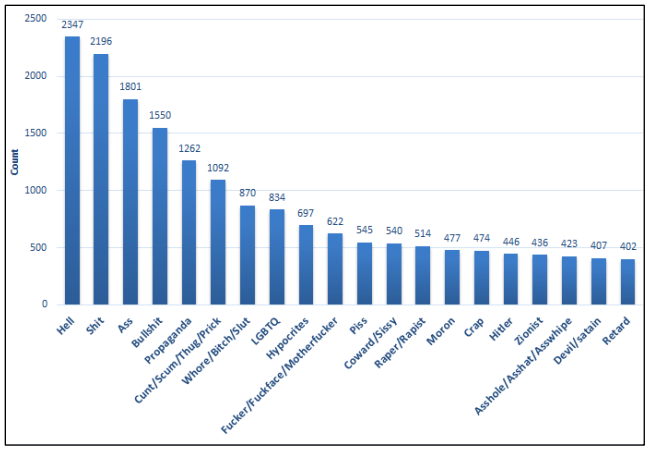

Figure 1: Distribution of Nouns Used in Offensive Context.

Regarding offensive verbs, most typical are Fuck/ Fucked/Fucking/Motherfucking (4,044 instances), Kill/Be Killed (2,983 instances), Raped (652), Hate (574), Stealing (273), and Screwed/Screwing (230). Additionally, Table 3 describes the most common offensive adjectives.

Table 3: Distribution of Offensive Adjectives in the Dataset.

Adjective	Frequency
Stupid	3,009
Disgusting	1,075
Pathetic	580
Ugly	330
Crappy/Shitty	326
Greedy	270
Retarded	229

To further explore the dataset, we run a topic model based on LDA (Latent Dirichlet allocation), as commonly done in computational social science. Since we already know that the targets are varied, we explore with three different number of topics (k=10, k=13, k=29). We find that the best number, in terms of interpretation, is 10 topics

[3] For example, http://www.bannedwordlist.com/lists/swearWords.txt
[4] Available at www.github.com/BLANK_FOR_REVIEW

(Table 4). When adding more topics, the results become difficult to interpret. This further encourages us to proceed with manual open coding, which uses human judgment for defining the categories.

Table 4: Topics from LDA Analysis, Named by Researchers.

Topic	Descriptive keywords
Race	white, black, racist, racism, race, blacks, hate, skin, color, american
Family	indi, girl, indian, animals, eat, year, animal, mother, baby, food
Police	police, cops, law, man, gun, guy, cop, shot, didn
Existence	don, way, really, good, say, world, right, time, need, life
Conspiracy	israel, money, world, country, land, government, oil, war, chin, live
Terrorism	muslims, muslim, islam, world, country, religion, isis, war, countries, terrorist
Politics	trump, americ, americans, president, country, obam, american, hillary, vote, clinton
Gender	women, men, woman, saudi, girls, man, arabi, culture, female, male
Globalization	basically, lol, japan, looks, kiss, bullying, water
Media	propaganda, aj, news, video, al, medi, qatar, anti, channel, western

Coding Guidelines

We applied open coding (Glaser & Strauss 1967), so that one of the authors coded the material until saturation was reached (i.e., no new categories emerged). During the coding process, categories were reorganized and added, and some subcategories merged into larger ones. This iterative nature of qualitative coding intends to improve the quality of the categories (Corbin and Strauss 1990).

The taxonomy was developed with the following guidelines in mind: 1) Read through the comments, identify themes and sub-themes. 2) While creating the categories, consider the hate target and the meaning of the comment. 3) When appropriate, apply hierarchy by first labeling the main theme, then a subtheme. 4) When classifying, include comments that are purposeful, i.e. intentionally hurtful. The last consideration was made because if hostility is not the purpose of the comment, it should not be classified as hateful. For example, "Trump is a bad president" is not hateful, but "Trump is an orange buffoon" is.

We considered linguistic attributes when annotating, as we are dealing with text. Swearing, aggressive comments, or mentioning the past political or ethnic conflicts in a non-constructive and harmful way, were classified as hateful. When there was uncertainty about an instance, it was discussed with other researchers to avoid a biased opinion. Finally, we utilized a coding dictionary so that, after identifying certain cue words for a category, such as "*Hitler [was right]*," we search the dataset with the cue words to find more observations for the corresponding category.

After saturation, two other researchers independently coded a random sample of 200 comments using the established taxonomy. We found substantial agreement (score = 75.3%) The agreement score was calculated by dividing the number of labels where two or more coders agreed by the number of possible values. A script was created to calculate this for each coded row, and the row-based agreements were averaged to get the overall agreement.

Taxonomy

The taxonomy has 13 main categories and 16 subcategories (29 in total). The main categories include targets and language, 9 describing targets and 4 the type of language. From our open coding, we find that hateful comments tend to focus both on groups of people (e.g., the Jews) and individuals. However, some hateful comments do not have a clear target (e.g., "Stupid people shouldn't breed"). Also, language may vary by target, so labeling both can be useful for modeling. Therefore, it makes sense to combine the type of language and the target of hateful comments. Figure 2 shows the hierarchical structure of the taxonomy, and Table 5 includes definitions and examples.

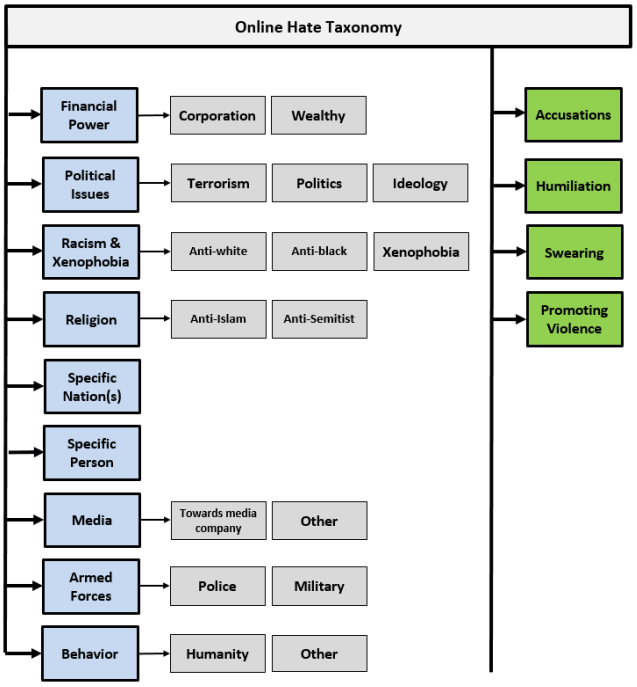

Figure 2: Hate Target Taxonomy. Hateful Language is in Green, Targets in Blue and Sub-targets in Grey Boxes.

The key distinctions of this taxonomy over previous work are: 1) it is more comprehensive, including 29 hate categories in total, which enables deeper understanding of online hate, and 2) it considers both hateful language and

targets, whereas previous work typically considers only one of the two.

Table 5: Taxonomy of Online Hate. Language Categories Highlighted with Thicker Borders; Others Are Targets.

Category	**Description**	**Examples**
Accusations	Accusing someone of something, without relevant evidence to support it. Accusations of lies, treason, all types of felonies, etc.	*"The refugees molest women and children, they shit in the swimming pool, and they worship some imaginary man in the sky who tells them to kill people."*
Promoting Violence	Calling people to deal with something using violence, asking for murders; threatening human life.	*"Disgusting cockroaches. Don't kick them. KILL them"*
Humiliation	Using words like: idiot, retard, stupid, dumb, trying to degrade someone.	*"I can't believe that we have so many ignorant, dumb, people in the US...though people in the US had brains"*
Swearing	Filthy language, bad words, swearing, non-polite.	*"fuck israel ... they will rot in hell ! even celebrities hate them"*
Financial Power	Hatred toward wealthy people and companies and their privileges. Pointing out their intentions to manipulate and commit crimes.	*"Lots of people getting rich off these pipelines...they could give a rats a_s about the people, the water, the land. Profit$ and me, me, me."*
Political Issues	Hate toward government, political parties and movements, war, terrorism, the flaws of the system.	*"That's how EU sees the freedom? Be naked, gay marriage, lesbian, celebrate dogs wedding. And thousand other bullshit things."*
Racism & Xenophobia	Racists comments toward black, white, asian. Generalizations about some characteristics, and hateful comments regarding refugees.	"*The white will always steal; FUCK YOU TO ALL WHITES RACIST*."
Religion	Everything about religion, including Judaism, Christianity, Islam, and religion in general. Both as a subject of hatred, or object.	*"FUCK the all the so call "Holy" Books. Bunch of grow ass adults believing in ancient book that was written by ancient people that think the Earth is flat and is in the center of the universe, those motherfucker deserve to die."*
Specific Nation(s)	Hate towards different countries, their systems, people (if the nationalities are mentioned), and certain events, like immigration, territory, and sovereignty.	*"Fucking Americans...I would go live in Mars if I could since it seems that Muslims might get persecuted in a couple of years like the Jews in nazi Germany"*
Specific Person	Hate toward specific people who can be regular people, politicians, millionaires, celebrities, or some other related to specific news.	*"He should be thrown back to the lions !Lets watch them eat him! Low life bastard no good sob Eye for and Eye"*
Media	Comments and emotional outbursts about bias and false statements made on purpose by the corrupted media.	*"Is there a news site THAT ISN'T BIASED?!? NO NEWS SITE IS UNBIASED ANYMORE."*
Armed Forces	Hate toward military, law enforcement, and the way they operate, which includes unethical behavior.	*"I hope that fucking cop burns in hell the man had head phones in fuck cops"*
Behavior	Hate toward the world, humanity, immoral actions of some part of the society, ignorant people, people that committed certain actions, and that have certain habits.	*"It's sad to see this but what's even worse is the people in the comment section making fun out of cows and etc. Stop comparing irrelevant things to this case. May God guide all of us especially those sick fuckers to the right path.*"

Model development

Overview

To achieve our research objective, we build two types of models: 1) binary classifiers that distinguish between hateful and non-hateful comments and 2) multiclass classifiers that provide granular information on hate targets and language. Since each comment in our data can belong to more than one category or subcategory of hate, we develop models that perform multilabel classification.

According to our taxonomy, there is a total of 29 categories to consider, including both main and sub-categories. In addition, there is one category for neutral comments, to make sure that the model is not biased toward detecting hate. Because some subcategories contain fewer than 10 labels, not enough for reliable classification, these subcategories were excluded when building the classifiers. The total number of categories and subcategories considered was therefore 21 out of the 29 in the taxonomy. Figure 3 shows the distribution of training data.

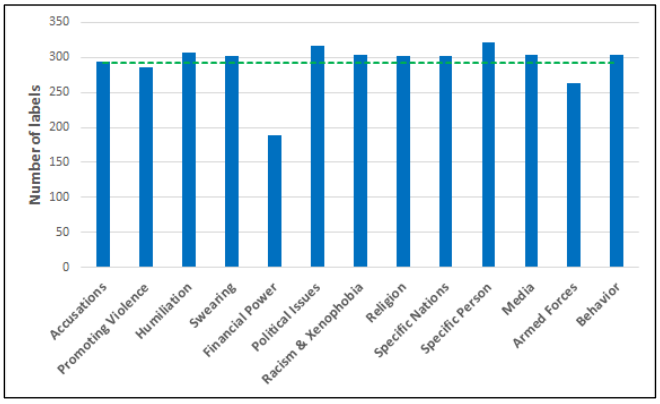

Figure 3: Training Data for the Main Categories. Green Line Indicates Average.

Except for financial power, classes are fairly balanced. We could not locate more samples for financial power in the research time frame. Non-hateful and hateful labels were used for the binary classification, while the labels shown in Figure 3 were used for the multilabel classification. Even after adding labels, the binary classes were unbalanced (hateful: 2,364; non-hateful: 1,357 labels), so we used Synthetic Minority Over-sampling Technique (SMOTE) to cope with this issue (Chawla et al., 2002).

To perform the supervised classification, three sets of feature categories were developed (cf. Nobata et al. 2016). The feature categories are n-gram, semantic and syntatic, and distributional semantic features. Before feature extraction, preprocessing was performed, removing stop words from comments and stripping the tokens of any trailing special characters or space. The preprocessing was not performed on the semantic features because special characters are essential for the feature computation.

N-gram Features

In this feature set, we used token n-grams that range between 1-3 grams. Each comment was split by space and the resulting tokens were used to generate the various n-gram features. For simplicity, we used the raw term frequency TF for the first set of n-gram features, and frequency–inverse document frequency (TF-IDF) for the second set of n-grams. TF-IDF captures the importance of a word to a document, in this case the collection of comments.

Semantic and Syntactic Features

Multiple semantic and syntactic features can be leveraged for classification purposes. For instance, hateful words and non-hateful (or polite) words can be considered as features that detect different types of hateful comments (Nobata et al. 2016). We used the following features:

- Count of exclamations, periods, question marks, punctuation, special characters, repeated punctuation, and quotes in each comment.
- Count of positive tokens; the list of positive words was from (Hu and Liu 2004) and Liu et al. (2005).
- Count of single-char. tokens in each comment.
- Count of the total number of discourse connectives in each comment (Pitler and Nenkova 2009). Here, we used a list of 100 discourse connectives from the Penn Discourse Treebank (PDTB)[5].
- Count of URLs in each comment.
- Length of the comment (in chars. and in tokens).
- Source of the comment (Facebook or YouTube).
- The average length of a token in each comment.
- Total number of capital letters in the tokens.
- Total number of emoticons in each comment.
- Total number of misspellings in each comment, comp. using the Enchant spell-checking library[6].
- Total number of modal words in each comment. The modal words that were used are: *can*, *could*, *may*, *might*, *must*, *will*, *would*, and *should*.
- Total number of tokens with non-alphabetic characters in the middle.
- Features based on a list of bad words[7]:
    - Checking if a comment contains a bad word.
    - If yes, count of bad words.
    - If yes, the ratio of bad words to all the tokens in the comment.

[5] https://www.seas.upenn.edu/~pdtb/
[6] https://www.abisource.com/projects/enchant/
[7] Here, we used several online sources, e.g. https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en

Distributional Semantic Features

Distributional word and text representations have garnered attention in the research community due to their success in solving a variety of machine learning problems. The idea behind these features is that lexical semantic aspects of the text are built using vector space models (Mikolov et al. 2013). Djuric et al. (2015) is one of the first works that tackled the problem of abusive language detection with word embeddings. There are different ways of using embeddings to represent text. Here, we rely on two basic techniques: word2vec and doc2vec. In word2vec, we use a pre-trained model constructed from Google's news dataset[8], which contains around 100 billion words. In this feature, we set the embedding vector to 300 dimensions.

In doc2vec features, the concept of words was extended to cover sentences, phrases, paragraphs, and documents (Le and Mikolov 2014). For this feature, we consider two types of text present in our dataset: the title of the YouTube video or Facebook post and the comment text. The approach to build doc2vec features is similar to Le and Mikolov's (2014) approach in which embeddings are trained using a skip-bigram model with a window size 10 and hierarchical softmax training. Like for the word2vec model, we set the embedding vector to 300 dimensions.

Experimental Evaluation

Next, we experiment with a set of machine learning algorithms to perform multi-label classification of the dataset. We give an overview of the classification performance on the labeled dataset using the harmonic mean score (F1) for the individual and combined features (see Tables 6 and 7).

Table 6: Binary Classification Results. Highest F1 Scores Bolded.

Feature / Classifier	TF	TF-IDF	Semantic	Word2vec	All feat.
Log. regression	**0.92**	**0.92**	0.77	0.66	0.91
Decision Tree	0.93	**0.94**	0.88	0.67	0.86
Random Forest	0.85	0.85	0.87	0.74	**0.89**
Adaboost	0.90	0.89	0.84	0.69	**0.92**
SVM	0.95	**0.96**	0.80	0.71	**0.96**

In this experiment, we used the 5,143 labels annotated using our taxonomy. This dataset was split into training and testing (33% for testing) the classification models. Five different models were used: Logistic Regression, Decision Tree, Random Forest, Adaboost, and Linear Support Vector Machine (SVM). For each model, we tuned the parameters using scikit-learn's[9] grid search method in Python. Moreover, we used pipelining to feed the features to the multilabel classifiers.

Table 7: Multilabel Classification Results. Highest F1 Scores Bolded.

Feature / Classifier	TF	TF-IDF	Sem.	Word-2vec	Doc2-vec	All feat.
Log. regression	**0.78**	0.77	0.02	0.36	0.36	0.04
Decision Tree	**0.77**	0.74	0.60	0.63	0.44	0.74
Random Forest	0.06	0.05	0.49	**0.66**	0.45	0.52
Adaboost	0.60	0.64	0.14	0.57	0.42	**0.70**
SVM	0.78	**0.79**	0.19	0.53	0.35	0.73

The average F1 score of the 21 categories and subcategories is used to report the overall performance as it combines both the precision and recall. For this analysis, we added the Doc2vec features. However, as Table 7 shows, n-gram features (TF and TF-IDF) produce the highest F1 score. Logistic Regression performs well with n-gram features, while the Random Forest performs poorly with the same features. It is also worth noting that the Decision Tree produces consistent results across all features.

The average precision of the best model, SVM, was 0.90, and recall 0.67. The recall score indicates our model is struggling to classify some categories, most notably religion (recall=0.3) and specific nations (0.43). In specific nations, there are many different country names, so the confusion is logical; also, the subcategories of religion are classified with a better recall (Judaism=0.76, Islam=0.75). In general, subcategories tended to perform better in the classification, probably because the language used is more precise. Considering this performance, we sum up the subcategory observations when analyzing the full dataset.

Analysis of Online Hate

We then apply the classifier to the full dataset to classify the comments by the main categories of hate. The results are displayed in Figure 4.

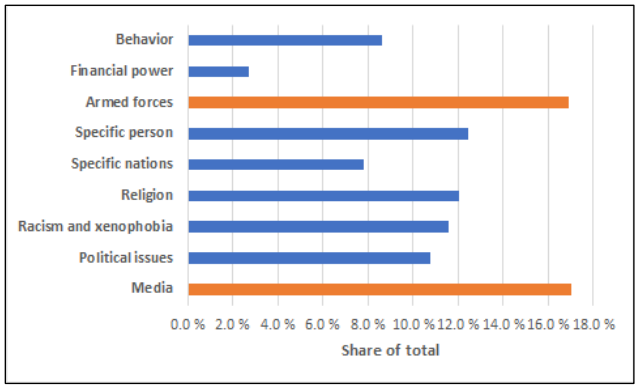

Figure 4: Analysis of Targets of Online Hate.

[8] https://code.google.com/archive/p/word2vec/
[9] http://scikit-learn.org/stable/

Most common is hate against Media (17.0% of instances), mostly against the particular news outlet which is referred to as "propaganda," "fake news," and "lies". Another popular target is Armed Forces (16.9%), particularly the police. The most typical language type is humiliation (31.5% of language observations) but swearing (29.3%) is also common. Somewhat alarming is the share of promoting violence (18.0%), clearly indicating that many comments are toxic. The reliability of prediction accuracy was verified by independent coding by one of the researchers of 200 randomly sampled comments. Calculating a simple agreement score lends support to the model's predictive accuracy (agreement with the model = 0.85).

Conclusion and Discussion

Mitigating online hate speech is important for reducing its harmful effects on the society, and this purpose represents one of the major impacts of computational social science on the society at large. The purpose of this research is to help community managers spot and remove malicious content by paving ways for automated or computer-aided moderation with the help of machine learning. Earlier efforts in this field have been myriad but tend to rely heavily on the use of a dictionary of hateful words, which has been found inadequate and rely on coarse categorizations providing little detailed information on the targets of hate.

To address these issues, we collected comments from a YouTube channel and Facebook page of a major online news organization from a given period and created training data by identifying hateful comments using human judgment, as encouraged by prior research (Sood et al. 2012a). We find that this annotation process is time-consuming but better captures the linguistical diversity of hateful comments than dictionary-based techniques. Using open coding that results in conceptually rich taxonomies, we found 13 main and 16 subcategories for online hate, whereas previous works typically identify only a few coarse categories. In addition to reporting accuracies, researchers should pay more attention to diverse categories of hateful comments, as they help understand the nature of hate in social media. Rich, inductive taxonomies capture both the linguistic diversity and the myriad targets for online hate.

Our main contribution is two-fold: first, the granular taxonomy of hateful online comments. Given the prevalence of toxic comments in everyday interactions in social media, such a taxonomy is needed. Thus, we identified categories and subcategories of hateful speech from the social media comments, forming a comprehensive taxonomy for machine learning. Second, we train a multiclass, multilabel model that classifies the hateful comments, experimenting with Logistic Regression, Decision Tree, Random Forest, Adaboost, and SVM. We found that SVM performed the best for the dataset (avg. F1 score = 0.79).

Applying the model to the full dataset, we find that media and the authorities (the police) are highly targeted among the commenters of the dataset. In conjunction, we found that surprisingly few comments were targeting other discussants. Most comments focused on outside targets – people were not arguing amongst themselves, but almost in isolation, or jointly at times, toward external targets. This indicates that isolated social media communities could become powerful catalysators of hate.

Our results can be explained in two ways: firstly, by considering recent media controversy and political polarization of people into different camps. And, secondly, by the research context: an online new media is likely to receive relatively more hate when reporting on political issues than when the topic is entertainment, for example. This suggests that the online hate could be different in other contexts, and that further research is needed to tie the topics of reportage with the types of hate in the related comments. Given these findings, especially the high degree of hate targeting news media, we encourage the media actors to collectively consider ways of defusing the hate rather than aggravating it. Especially in the current climate of polarization, it is likely that politically loaded news drives more hateful speech and dissonance than it relieves it.

While more research is needed to validate and extend our work beyond the chosen context, we have demonstrated the applicability of automatic classification of online hate at a granular level. Future work can improve upon this research by increasing features; e.g., while conducting the research, YouTube made new information available in the JSON output. These features could be useful for improving model accuracy. Also, experimental studies controlling the hate in comments are rare. Future studies could filter hate speech in real systems and analyze the impact on user perceptions. For example, Salminen et al. (2018) found that hateful comments have the potential to contaminate user perceptions toward automatically generated personas.

Finally, we observe the following challenges for automatic detection of online hate speech:

- **Interpretation problem** – the interpretation and the intensity of perceptions of hate speech may differ among individuals.
- **Linguistic variety** – the language used for hate speech is in a state of constant flux, stressing the importance of concept drift and "living models".
- **Danger of over-moderation** – models should detect comments that are actually harmful, without jeopardizing freedom of speech.
- **Limits of automation** – hate can be seen reflective to individuals not feeling well, an issue which technology has only limited ability to solve, requiring social solutions and human supervision.

References

Akerlof GA (1970) The Market for 'Lemons': Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics* 84(3): 488–500.

Badjatiya P, Gupta S, Gupta M, et al. (2017) Deep Learning for Hate Speech Detection in Tweets. In: *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, pp. 759–760.

Chawla NV, Bowyer KW, Hall LO, et al. (2002) SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16: 321–357.

Corbin JM and Strauss A (1990) Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13(1): 3–21.

Davidson T, Warmsley D, Macy M, et al. (2017) Automated Hate Speech Detection and the Problem of Offensive Language. In: *Proceedings of Eleventh International AAAI Conference on Web and Social Media*, Québec, Canada.

Del Vicario M, Vivaldo G, Bessi A, et al. (2016) Echo Chambers: Emotional Contagion and Group Polarization on Facebook. *Scientific Reports* 6: 37825.

Djuric N, Zhou J, Morris R, et al. (2015) Hate Speech Detection with Comment Embeddings. In: *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, New York, NY, USA: ACM, pp. 29–30.

Glaser BG and Strauss AL (1967) *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Transaction Publishers.

Hosseini H, Kannan S, Zhang B, et al. (2017) Deceiving Google's Perspective API Built for Detecting Toxic Comments. *arXiv:1702.08138 [cs]*. Available from: http://arxiv.org/abs/1702.08138 (accessed 7 January 2018).

Hu M and Liu B (2004) Mining and summarizing customer reviews. In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 168–177.

Hughey MW and Daniels J (2013) Racist comments at online news sites: a methodological dilemma for discourse analysis. *Media, Culture & Society* 35(3): 332–347.

Kramer ADI, Guillory JE and Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111(24): 8788–8790.

Kwon KH and Gruzd A (2017) Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research* 27(4): 991–1010.

Le Q and Mikolov T (2014) Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1188–1196.

Liu B, Hu M and Cheng J (2005) Opinion observer: analyzing and comparing opinions on the web. In: *Proceedings of the 14th international conference on World Wide Web*, ACM, pp. 342–351.

Mikolov T, Chen K, Corrado G, et al. (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mondal M, Silva LA and Benevenuto F (2017) A Measurement Study of Hate Speech in Social Media. In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, New York, NY, USA: ACM, pp. 85–94.

Nobata C, Tetreault J, Thomas A, et al. (2016) Abusive Language Detection in Online User Content. In: *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, pp. 145–153.

Park JH and Fung P (2017) One-step and Two-step Classification for Abusive Language Detection on Twitter. *arXiv preprint arXiv:1706.01206*.

Pfeffer J, Zorbach T and Carley KM (2014) Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications* 20(1–2): 117–128.

Pitler E and Nenkova A (2009) Using syntax to disambiguate explicit discourse connectives in text. In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Association for Computational Linguistics, pp. 13–16.

Rajadesingan A, Zafarani R and Liu H (2015) Sarcasm detection on twitter: A behavioral modeling approach. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ACM, pp. 97–106.

Saleem HM, Dillon KP, Benesch S, et al. (2017) A Web of Hate: Tackling Hateful Speech in Online Social Spaces. *arXiv:1709.10159 [cs]*. Available from: http://arxiv.org/abs/1709.10159 (accessed 3 October 2017).

Salminen J, Nielsen L, Jung S-G, et al. (2018) "Is More Better?": Impact of Multiple Photos on Perception of Persona Profiles. In: *Proceedings of ACM CHI Conference on Human Factors in Computing Systems (CHI2018)*, Montréal, Canada.

Silva L, Mondal M, Correa D, et al. (2016) Analyzing the Targets of Hate in Online Social Media. In: *Proceedings of Tenth International AAAI Conference on Web and Social Media*, Palo Alto, California.

Sood S, Antin J and Churchill E (2012a) Profanity Use in Online Communities. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, New York, NY, USA: ACM, pp. 1481–1490.

Sood S, Churchill EF and Antin J (2012b) Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology* 63(2): 270–285.

Walker S (1994) *Hate speech: The history of an American controversy*. U of Nebraska Press.

Wright L, Ruths D, Dillon KP, et al. (2017) Vectors for Counter-speech on Twitter. In: *Proceedings of the First Workshop on Abusive Language Online*, pp. 57–62.