Hate Speech Detection *WiSe 23-24*
Classical ML Methods + Features

Image taken from https://deepsense.ai/artificial-intelligence-hate-speech/

(Dr.) Özge Alaçam

Computational Linguistics

oezge.alacam@lmu.de

Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)

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

- Salminen, J., Almerekhi, H., Milenković, M., Jung, S. G., An, J., Kwak, H., & Jansen, B. (2018, June). Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 12, No. 1).

the Generalization Problem

- There is a lack of methods of understanding the types and targets of hate speech
- 1. How can hateful comments on social media be automatically detected and classified?
- 2. What are the common targets of online hate speech?

A research with classical ML methods

- manually labeled **5,143** hateful expressions posted to YouTube and Facebook videos among a dataset of **137,098** comments from online news media.

- created ***a granular taxonomy*** of different types and targets of online hate
  - *open coding technique (?)*

- Trained ***classical machine learning models*** to automatically detect and classify the hateful comments in the full dataset.
  - Logistic Regression, Decision Tree, Random Forest, Adaboost, and Linear SVM

  - The task is to detect and categorize hateful comments in the context of online news media.

Motivation

- Dictionary-based methods are powerful indicators for hateful comments,
- But they are not enough to detect all variants of hate speech (Saleem et al., 2017)
  - ❖ False positives, such as: “I really hate owing people favors,”
  - ❖ “**** people” >>> “people” can also be in the hate lexicon.
  - ❖ “I hate police officers.” but miss “police officers are dogs.”
- Keywords are also prone to missing sarcasm and forms of humor, (Rajadesingan et al. 2015).
- The blacklist (a special collection of hateful words and insults) requires constant updates (Nobata et al., 2016)
- >>> more granular models are needed,

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)

1st step: Exploration using a lexicon

- A public sources of hateful words (200 selected)
  - 22,514 comments (16.4%) contain these hateful wordings

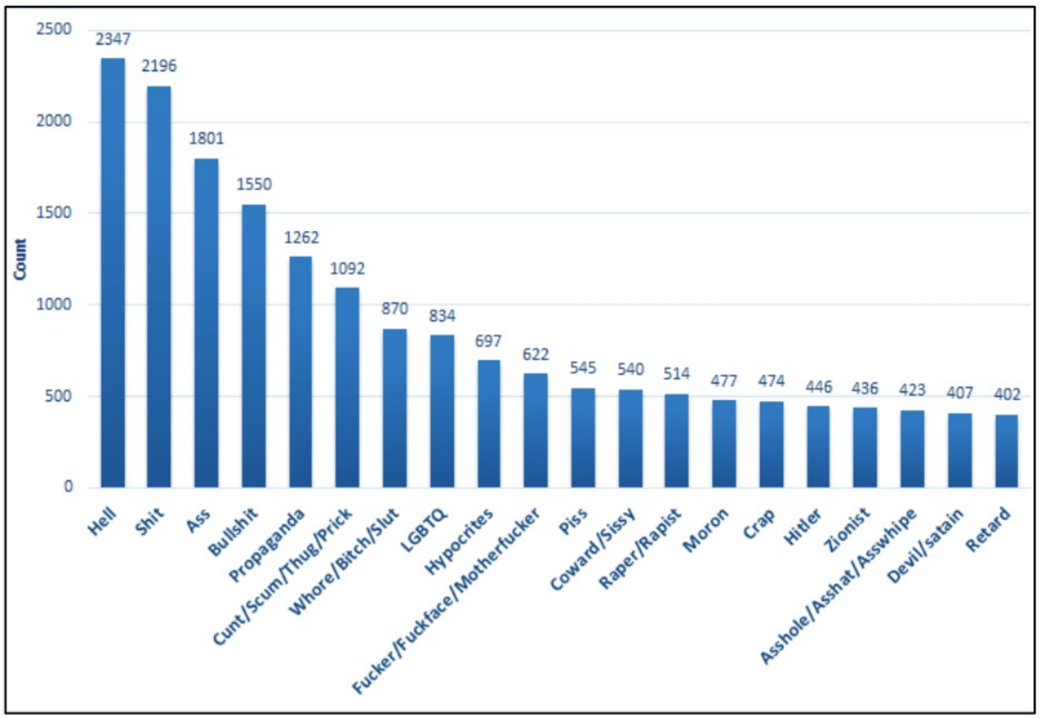


Figure 1: Distribution of Nouns Used in Offensive Context.

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

2nd step : Find clusters of targets using LDA

- a topic model based on LDA (Latent Dirichlet allocation)

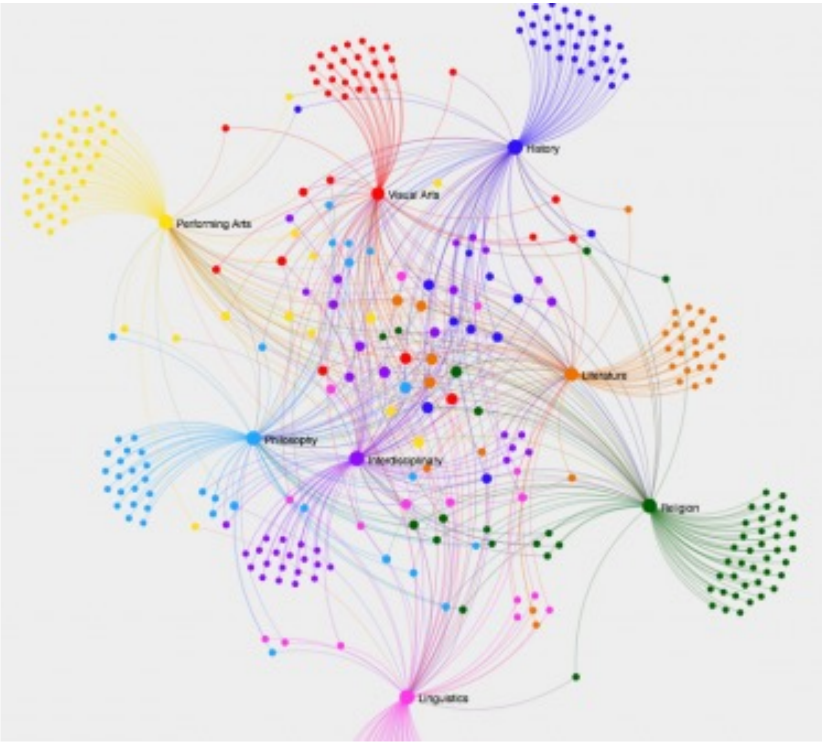
- Three different number of topics (targets) (k=10, k=13, k=29).

  find the best k (in this case it is 10)

Topic Modeling Algorithms

- latent dirichlet allocation (LDA)
- latent semantic analysis (LSA)
- probabilistic latent semantic analysis (PLSA)

label the documents with "unobserved" topics by classifying the words!



Source: *Vajjala et al. (2020) Practical Natural Language Processing, CH7*

Topic Modeling

Document-term matrix: *5 X 6*

	W1	W2	W3	W4	W5	W6
D1	0	3	0	0	1	2
D2	1	0	0	1	1	1
D3	2	1	2	2	4	2
D4	1	1	1	4	0	0
D5	0	1	2	1	0	4

Factorized Matrices: *5 X 4 document – topic matrix* & *4 X 6 topic – term matrix*

	K1	K2	K3	K4
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
D4	1	0	1	0
D5	0	1	1	1

	W1	W2	W3	W4	W5	W6
K1	1	0	0	1	0	0
K2	0	1	1	0	1	1
K3	1	1	0	1	1	0
K4	1	0	0	0	1	0

! check the implementation

Topic Modeling (Example)

- A collection of 5 documents, that contain a single sentence;

	Sentences	Topic-A	Topic-B
D1	I like to eat broccoli and bananas.		
D2	I ate a banana and salad for breakfast.		
D3	Puppies and kittens are cute.		
D4	My sister adopted a kitten yesterday.		
D5	Look at this cute hamster munching on a piece of broccoli.		

Example taken from Vajjala et al. (2020) Practical Natural Language Processing, CH7

Topic Modeling - Example (cont.`)

- Learning a topic model on this collection using LDA :

A topic model only gives a collection of keywords per topic.

- **Topic A:** broccoli, bananas, breakfast, munching
- **Topic B:** puppies, kittens, cute, hamster

	Sentences	**Topic-A**	**Topic-B**
D1	I like to eat broccoli and bananas.	100%	
D2	I ate a banana and salad for breakfast.	100%	
D3	Puppies and kittens are cute.		100%
D4	My sister adopted a kitten yesterday.		100%
D5	Look at this cute hamster munching on a piece of broccoli.	60%	40%

Back to the study…

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

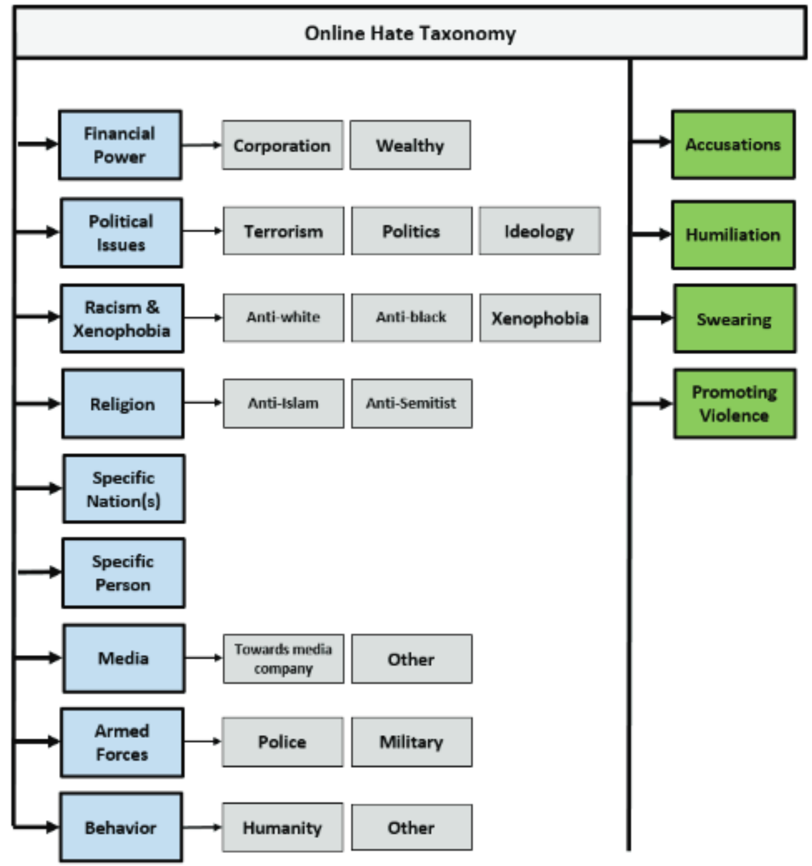


Figure 2: Hate Target Taxonomy. Hateful Language is in Green, Targets in Blue and Sub-targets in Grey Boxes.

Step3: Deciding on the classification task

1. binary classifiers that distinguish between hateful and non-hateful comments

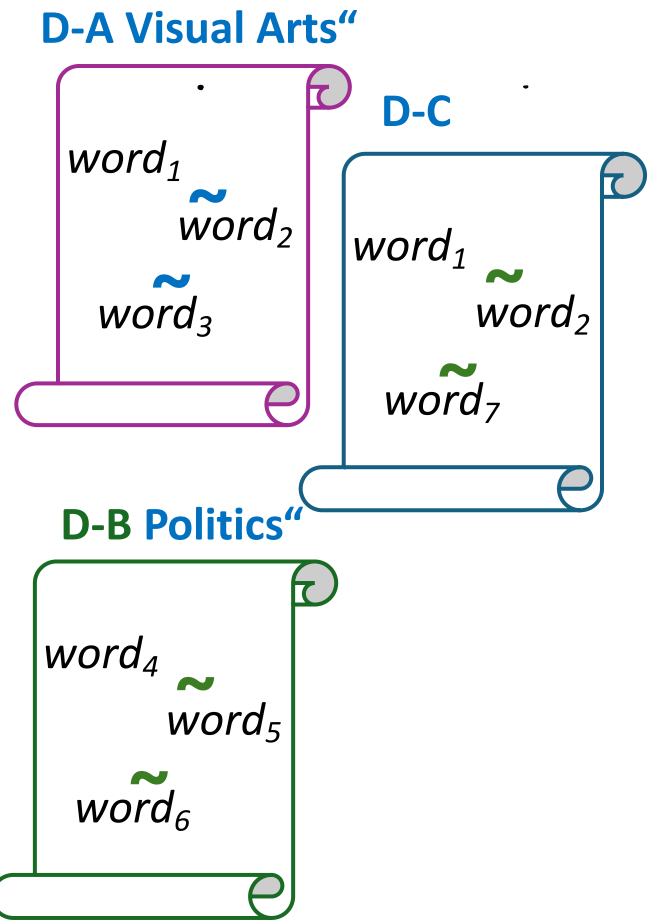
2. Multiclass classifiers that provide granular information on hate targets and language

Step4: Deciding on the features

- n-gram features
  - n-grams that range between 1-3 grams.
  - Created using term frequency (TF) and frequency-inverse document frequency (TF-IDF)
- Semantic and syntactic features
- Distributional semantic features

Brief Recap:
Vector Representations &
TF-IDF Vectorizer

Vector representations



- Words that occur in similar contexts tend to have similar meanings.
  - Excluding frequent words
- So similar words have similar vectors.
- Two documents that are similar will tend to have similar words

	D-A	D-B
Word1	**1**	0
word2	**1**	0
word3	1	0
word4	0	1
word5	0	1
word6	0	1
word7	0	0

Vector representations

- Frequency based models
  - Tf-idf models, PMI
- Static word embeddings;
  - word2vec, GloVe, fasttext
- Deep contextualized representations
  - ELMo, BERT, GPT, Llama etc.

Vector representations

- simple frequency (count vectorizer) isn't the best measure of association between words.
  - raw frequency is very skewed
  - not very discriminative (the, a, did, they etc.)
- Words that occur nearby frequently are more important than words that only appear once or twice.
  - word association

Tf-idf Model

- Baseline model
- The meaning of a word is defined by a simple function of the counts of nearby words.
- TF (term frequency) measures how often a term or word occurs in a given document.
- IDF (inverse document frequency) measures the importance of the term across a corpus.
- Note: It produces very long vectors that are sparse,

    i.e. contain mostly zeros (each word is represented with a dimension)

- Back To Study >> They use tf and tf-idf vectorizers to extract the n-gram features

Semantic And Syntactic 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-character tokens in each comment.
- Count of the total number of discourse connectives in each comment (Pitler and Nenkova 2009.
- Count of URLs in each comment.
- Length of the comment (in chars. and in tokens).

Semantic And Syntactic Features

- 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
- Total number of modal words in each comment.
- Total number of tokens with non-alphabetic characters in the middle.

Distributional Semantic Features

- Word2vec: pre-trained model constructed from Google's news dataset, which contains around 100 billion words (300 dimensional)

- Doc2vec: embeddings trained using a skip-bigram model with a window size of 10 and hierarchical softmax training (300 dimensional).

- The textual input:
  - the title of the YouTube video or Facebook post
  - the comment text.

Step 5: Experimental Evaluation on 5K comments

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

Feature / Classifier	**TF**	**TF-IDF**	**Semantic**	**Word2vec**	**All feat.**
Log. regression					
Decision Tree					
Random Forest					
Adaboost					
SVM					

Step-6: Understanding the results

- Evaluation on full dataset (137K comments) with the selected model (SVM))

Language type:

- the most typical language type is humiliation (31.5% of language observations)
- Next is swearing (29.3%)
- Then promoting violence (18.0%)

Target Analysis:

- Most frequent target is Media, Armed Forces,
- Specific Forces

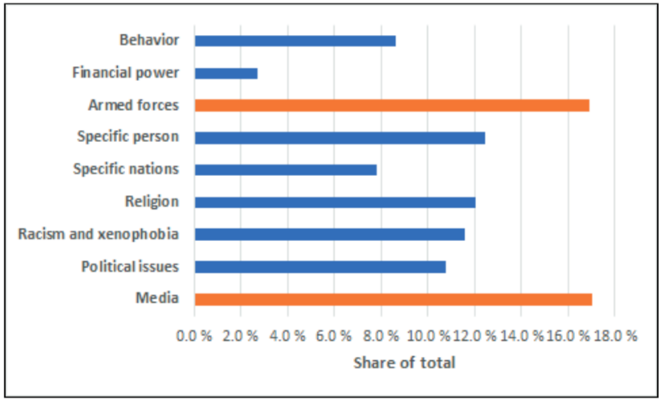


Figure 4: Analysis of Targets of Online Hate.