Hate Speech Detection with Bayesian Networks

Susana Benavidez, Pratyusha Javangula , Andrew Lapastora, Derek McCreight; *Mentor: Chuma Kabaghe*

Introduction

- Hate speech and offensive language threaten the health and longevity of online communities, and can have long-lasting adverse effects on those targeted.
- Still, these two are not identical, and automated hate speech detection systems must be able to distinguish the two.
- We present the results of several linear classifiers and a Bayesian network applied to the task of hate speech detection of Tweets.
- We applied a Bayesian network to this task to discover the relationships between features of the Tweets and their effect on whether a Tweet is most likely a posteriori to be hate speech, offensive language, or neither.

Dataset

- We used the 3-class dataset of Tweets collected by Davidson et al. (2017)[1].
- Labels were generated via a crowdsourcing website. Between 3-6 raters classified each tweet, with mean inter-annotator agreement of 92%
- Possible labels: "Hate speech", "Offensive language", "neither"

Example Tweets and Corresponding Labels

1. *Hate speech: "Think it's okay to take my property and break it? F*** you b****"*
2. *Offensive language: "I'm bringing booty back. Go ahead and tell them skinny b****** that."*
3. *Neither: "This 8 yr old on #masterchefjunior made my chicken look like trash.*

Bayesian Network

- Structure learning with score-based evaluation + greedy/Chow-Liu tree search algorithm
- Parameter learning with loopy belief propagation (belief at a node in the graph corresponds to marginal distribution over that variable/feature)
- Classification as inference over likely value of label given evidence (in this case, variables corresponding to features extracted from tweet)

[1]Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speechdetection and the problem of offensive language, 2017b.

Class label	Mean unigrams	Mean bigrams	Mean characters
Hate speech	15.89	14.89	4.34
Offensive language	16.11	15.11	4.04
Neither	17.45	16.45	4.43

Table 1: Text statistics in dataset.

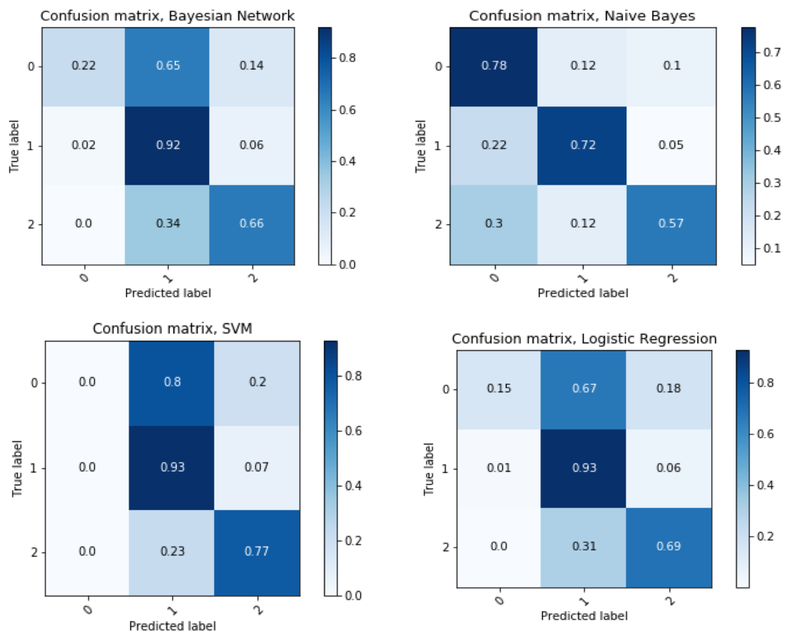
Class	Number of samples
Hate speech	1430
Offensive language	19190
Neither	4163

Table 2: Classes and their count in dataset.

Features

- Employed a variety of linguistic and semantic features to take into account language complexity
- We leveraged Hatebase-derived contextual features for tweets which reference characteristics such as class, disability, gender, nationality, religion, sexual orientation, or if it is profane/racist
- We discretize by comparing the count of a feature in a given tweet to the count of the feature over the entire dataset at different quantiles

Results



Evaluation

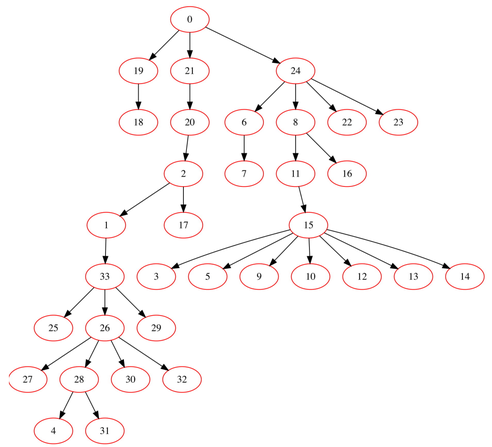
- We used the F1 macro score with a particular focus on recall for hate speech and offensive language
- Zero rule baseline that always predicts majority class
- Human performance oracle on 100 held out samples
- Other baselines: logistic regression, SVM with RBF kernels, and Naive Bayes

Class	Precision	Recall	F1 score
Hate Speech	1.00	0.33	0.50
Offensive language	0.88	0.49	0.63
Neither	0.30	0.94	0.46
Macro-average	0.73	0.59	0.53

Table 3: Precision, recall, and F1 scores for each class in the dataset as well as macro-average, obtained from human performance oracle.

Class	Precision	Recall	F1 score
Hate Speech	0.00	0.00	0.00
Offensive language	0.79	1.00	0.88
Neither	0.00	0.00	0.00
Macro-average	0.26	0.33	0.29

Table 4: Precision, recall, and F1 scores for each class in the development dataset as well as macro-average, obtained from the zero rule baseline.



Features	Nodes
Word Counts	0
Swear Counts	1
Profanity Counts	2
Contains Profanity	3
Racist Counts	4
Contains Racist	5
Mention Counts	6
Contains Mention	7
Contains Hashtag	8
Hashtag Counts	9
Emoji Counts	10
Contains Emoji	11
Negative Seniment Emoji Score	12
Positive Sentiment Emoji Score	13
Neutral Sentiment Emoji Score	14
Overall Sentiment Emoji Score	15
Consecutive Punctuation	16
Majority Uppercase Letters	17
First Word Second or Third Person Pronoun	18
More Second or Third Person Pronouns than First	19
Majority Negative Words	20
No Positive Words	21
Contains URL	22
Contains Non Alphanumeric Word	23
Misspelling Count	24
Judgement Disagreements	25
About Gender (Hatebase)	26
About Religion (Hatebase)	27
About Ethnicity (Hatebase)	28
About Sexual Orientation (Hatebase)	29
About Disability (Hatebase)	30
About Class (Hatebase)	31
About Nationality (Hatebase)	32

Table 5: Network Features and respective node label of Chow-Liu structure graph

Conclusion & Future Work

- Incorporate continuous features rather than only discretized features.
- Employ a larger dataset for our task
- Utilize more computational power to provide for a more extensive structure search and more variables in the network