Paper 5204-2020

Hate Speech Classification of social media posts using Text Analysis and Machine Learning

Venkateshwarlu Konduri, Sarada Padathula, Asish Pamu and Sravani Sigadam,
Oklahoma State University

ABSTRACT

Hate crimes are on the rise in the United States and other parts of the world. Hate speech is one tool that a person or group uses to let out feelings of bias, hatred and prejudice towards a religion, race, ethnicity, ancestry, sexual orientation, gender or disability thereby spreading hatred. This paper focuses on how SAS® Enterprise Miner's Text Analytics was used to develop a model that categorizes tweets based on their content, specifically hateful vs normal. After sampling and cleaning of the data and breaking the tweets down into quantifiable components, different models were built and compared. The best performing model was used to score unseen data, achieving reasonable accuracy in classification. This paper touches upon how text analytics could be harnessed by organizations like Twitter for encouraging civic responsibility in its users. By providing a feature at the user-level which allows tweets to be labelled as a particular category as they are typed, the users might be given an opportunity to review and possibly modify any hateful tweets before posting them.

INTRODUCTION

Social media has revolutionized the way people stay connected, exchange ideas, share information and voice opinions. Out of the total world population of 7.7 Billion people, approximately 3.5 Billion people use various social media platforms. As the number of people adopting social media usage increases, the use of these media to vent hatred also seems to increase.

Twitter is a common medium for voicing opinions, producing up to 500 million tweets per day. While on one side tweets promoting social good can be seen, on the other side tweets with extreme views, aggression, threat and hatred can also be seen. Sometimes users might be unaware of their tweet being perceived as hateful and of the consequences of tweeting a such a tweet. There are instances where people have been arrested or detained for allegedly sending "offensive" messages via social media. Social media companies have put in place rules, policies and algorithms to restrict what should and should not be posted to make them a safer and less toxic place. They remove abusive/hate content, block accounts and take legal actions against users in violation of the rules, usually after such tweets are posted and reported.

PROBLEM STATEMENT

Reports indicate that arrests for aggressive, threatening or hateful speech on social media are not uncommon and some of them are of people who unintentionally express their negative feelings or whose feelings were negatively interpreted[1]. For example, as reported by a Business Insider article, a 16-year-old teen on his way to a Pink concert got a little too excited about the artist performing her song, "Timebomb.", tweeted something that sounded suspicious, when taken out of context: "@Pink I'm ready with my Bomb. Time to blow up

[1] https://www.businessinsider.com/tweets-that-got-people-arrested-2013-7

#RodLaverArena. B----.". He was immediately arrested and not released until the context of the tweet was clarified.

This is just one instance of an array of such incidents reported. A good way to stay out of such situations is probably to be cautious about what to tweet and having a mechanism to flag potentially offensive content. The aim of this project is to aid social media users rethink about the content they are about to post, so that any unintended consequences might be avoided. This paper's focus is on building predictive models for classifying text data as hate speech or normal speech. The results obtained from scoring the model, can be used to develop an extension or an application which can provide "offensiveness" or "hatefulness" scores on a scale (for e.g., of 0-5) to the user based on the tweet content. Using this information, a user can make a decision on whether to post their tweet or not (and avoid any potential legal consequences). Apart from providing this feature to users who wish to make use of it as a precaution for not tweeting something hateful in the first place, Twitter could also use it for automatic flagging of potentially problematic/hateful tweets for internal review.

METHODS

DATA COLLECTION, PREPARATION AND VALIDATION

Two datasets were used for this paper. The first one was the Tweets data set in CSV file format with tab separated columns obtained from the research paper[2] cited below. This original Tweets CSV file (with approximately 100,000 observations) which was used for training the model, had three columns namely Text (the actual tweet), Label (label given to each tweet based on judgement of multiple annotators: normal, hateful, abusive and spam) and Nvotes (number of crowdsourced annotators who agreed with the Label given to the tweet). The second dataset which was used for scoring the model was another Twitter dataset in CSV file format with tab separated columns collected from GitHub[3]. This dataset (with approximately 24,784 observations) had six columns namely Count, hate speech, offensive language, neither, class (annotated as one of: normal, hateful and offensive) and tweet. This dataset consisted of 5,593 records with 1,430 labelled as "hate" and 4,163 labelled as "normal" tweets.

Both the CSV files were imported using SAS® Enterprise Guide import option to create SAS datasets. Only the columns considered necessary i.e. text and label of both datasets were used for modelling and analysis. Since the focus of this research paper is to analyze only hate related tweets, observations from the training dataset labelled "spam" and "abusive" were removed and a new dataset having approximately 59,000 observations was created. Similarly, URLs and Twitter usernames were removed. This new dataset was then imported to SAS® Enterprise Miner. The dataset was found to be imbalanced with respect to the Label column, as it had only 8% of observations with hateful label. So, a stratified sampling method was used to create a balanced data set. Next, the data was partitioned into 70% training and 30% validation which were used for modelling and analysis.

PROJECT APPROACH

Figures A.(i) and A.(ii) in Appendix-A outline the project approach.

DATA CLEANING, MANIPULATION AND RATIONALE

After the data partitioning, Text parsing node was used on the text column. The parsing process cleaned and modified the tweets' textual data by converting all the text to lowercase.

[2] https://github.com/ENCASEH2020/hatespeech-twitter

[3] https://github.com/t-davidson/hate-speech-and-offensive-language/tree/master/data

Then the text was tokenized, meaning entire text was split into individual words to form separate variables; a number was assigned to each word representing the number of times the word was used in that tweet (i.e. frequency). Stop words, numbers, punctuation marks, non-English words and characters like â were removed. SAS® Enterprise Miner entity recognizer was used to remove names of entities like persons, Organizations, Currency symbols, Addresses, Locations and phone numbers.

After the text was parsed, it was filtered to reduce the number of words. Importance of a term/word to a piece of text is defined by Term Frequency Inverse Document Frequency (TF-IDF). If a word did not appear in at least four tweets or if a word was not important to the text (based on TF-IDF), it was removed. Next, Text topics were created. The purpose of Text topics was to identify a collection of similar pattern of words existing in multiple tweets and then form topics by grouping those words. Text topics created in this process have many-to-many relationships to tweets, i.e. one tweet can be linked with many topics. Text topic and Text clustering are similar mechanisms that SAS® Enterprise Miner provides for grouping words into common themes, but for the context of this paper, the former was considered more applicable. Text topics can associate a tweet to multiple groups unlike Text clustering which can link each tweet to only one group. For example, if a tweet contained both the 'Happy and Morning', it could have fallen into two different text topics, both topics highly correlated with the text labelled as normal. Whereas, if text clustering were used, it would have been assigned to only one cluster. Raw text topic probabilities were used in the predictive models, which meant that each document had a probability of belonging to each of the 25 generated text topics. If a tweet expressing anger uses words associated with normal labelled text such as 'today' or 'you' as well as words associated with hate labelled texts such as 'f***' (an expletive), the model factors the use of the terms in both labelled topics when predicting the hatefulness of tweet.

RESULTS

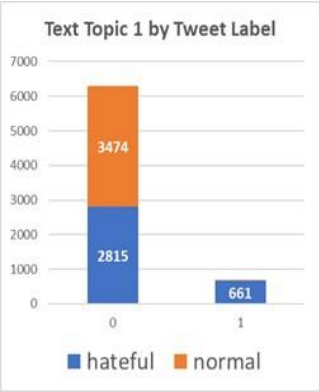
TEXT TOPICS

25 text topics were created and could sufficiently distinguish between hateful and normal terms in the text corpus. The text topics were able to differentiate between topics related not only to racial hate, but also religious hate and hate targeting sexual orientation based on the words commonly used in the tweets. This is because each type of hate has distinct words being used.

Example:

Racial hate related Topic:

Key terms: N****, Retard, mad



Normal Topic:

Key Terms: Happy, Good, Love

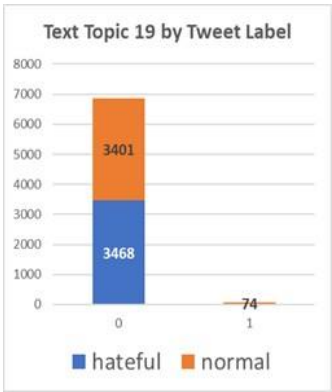


Figure 1: Text Topics for Hate and Normal speech

MODELS

Following is a summary about the models built and results generated:

- Predictive models were built using the 25 text topics. Logistic Regression, Random Forest, Decision Tree and Gradient Boosting modelling techniques were used.
- These modelling techniques were chosen because all the four models can either: handle data with higher dimensionality with relative ease; or easy interpretability of results.
- The Gradient boost tree model with validation misclassification rate of 18% was selected as champion model (See Appendix B for model comparison results). The champion model has a sensitivity rate of 90% (i.e. the ratio of correctly classified hate tweets to total number of actual hate tweet is 90%).

SCORING RESULTS

Scoring was performed on the second dataset. The misclassification rate on this scoring dataset was 20%. The results of confusion matrix are shown below:

	Actual: Hate	Actual: Normal
Predict: Hate	949 (16.96%)	621 (11.10%)
Predict: Normal	481 (8.60%)	3542 (63.33%)

Table 1. Confusion Matrix on the scoring data

GENERALIZATIONS

The accuracy of the champion model in the classification of the scoring and validation datasets were 80.2% and 81.64% respectively; the Sensitivity ratios were 66% and 90% respectively. The scoring results were consistent with validation results and the model was able to categorize hate words into different categories like racial words, sexually abusive words, threatening words, and normal words as normal.

FUTURE PROSPECTS

In the future iterations, Sentiment analysis and the use of attributes of tweets such as number of retweets, likes and comments may be used to enhance the predictive power and accuracy of the model. Deep Learning models like Recurrent Neural Network (RNN) and 1D Convolution Neural Networks (CNN1D) can also be leveraged to identify underlying patterns and flag hate speech. This analysis can also be extended to platforms such as Facebook, Instagram, blogs and articles on the Internet.

CONCLUSION

This research paper demonstrated an approach to classify speech as hateful vs normal based on the content using methodologies like Text analysis along with Machine Learning. Accuracy of the model was above 80% for both validation and scoring datasets and this can further be improved by leveraging the attributes and sentiment of tweets. The model can be deployed

through browser extensions of app features to better assist a social media user in avoiding posting hateful content.

REFERENCES

10 Ways to Get Arrested for Tweeting by Alyson Shontell. "Business Insider" - July 18, 2013. Available at https://www.businessinsider.in/law-order/10-ways-to-get-arrested-for-tweeting/articleshow/21134994.cms.

Hate Speech on Social Media: Global Comparisons by Zachary Laub." Council on Foreign Relations" - June 7, 2019. Available at https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

Hate and Abusive Speech on Twitter. GitHub. Available at https://github.com/ENCASEH2020/hatespeech-twitter

Founta, Antigoni-Maria and Djouvas, Constantinos and Chatzakou, Despoina and Leontiadis, Ilias and Blackburn, Jeremy and Stringhini, Gianluca and Vakali, Athena and Sirivianos, Michael and Kourtellis, Nicolas." Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior". June 2018. 11th International Conference on Web and Social Media, ICWSM 2018, AAAI Press. Available at https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17909/17041

John T. Nockleby. Hate Speech, pages 1277–1279. Macmillan, New York, 2000.

Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS® Dr. Goutam Chakraborty, Murali Pagolu, Satish Garla

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Venkateshwarlu Konduri
Oklahoma State University
venateshwarlu.konduri@okstate.edu

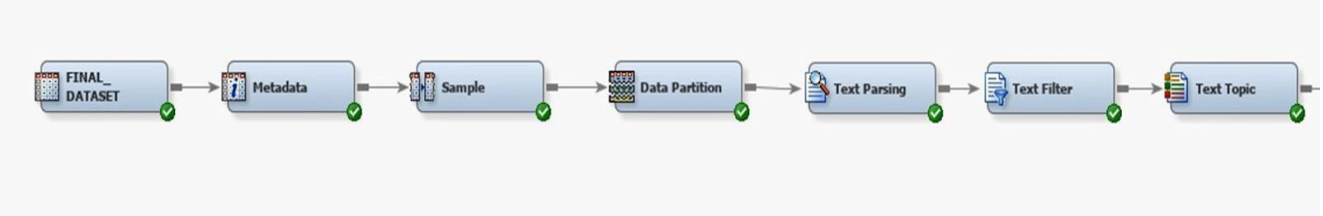
Sarada Padathula
Oklahoma State University
sarada.padathula@okstate.edu

Asish Pamu
Oklahoma State University
ashish.pamu@okstate.edu

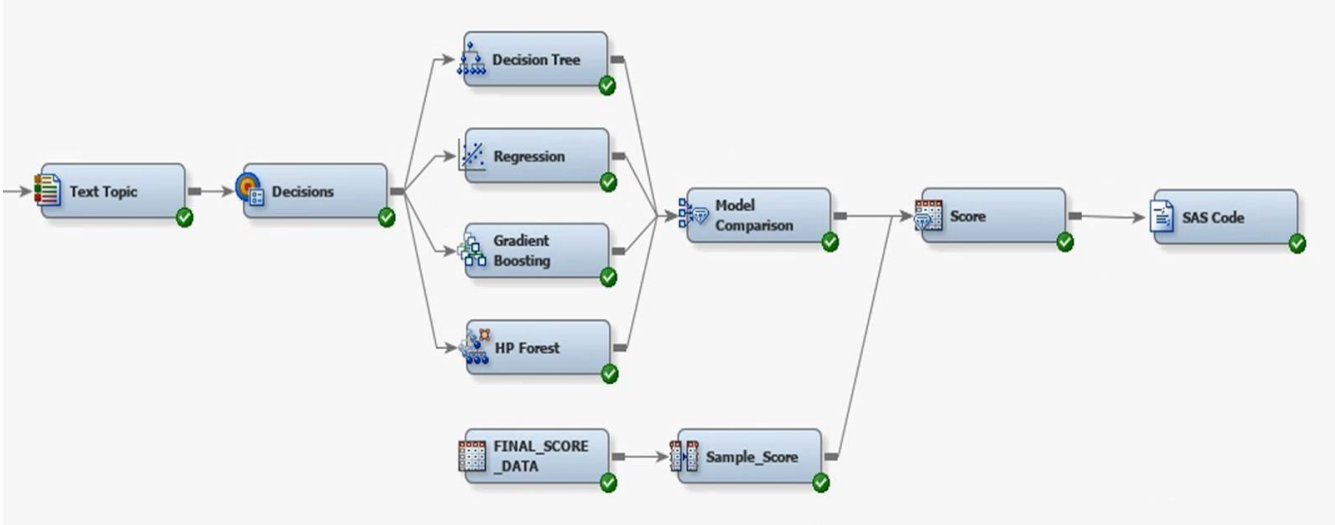
Sravani Sigadam
Oklahoma State University
sravani.sigadam@okstate.edu

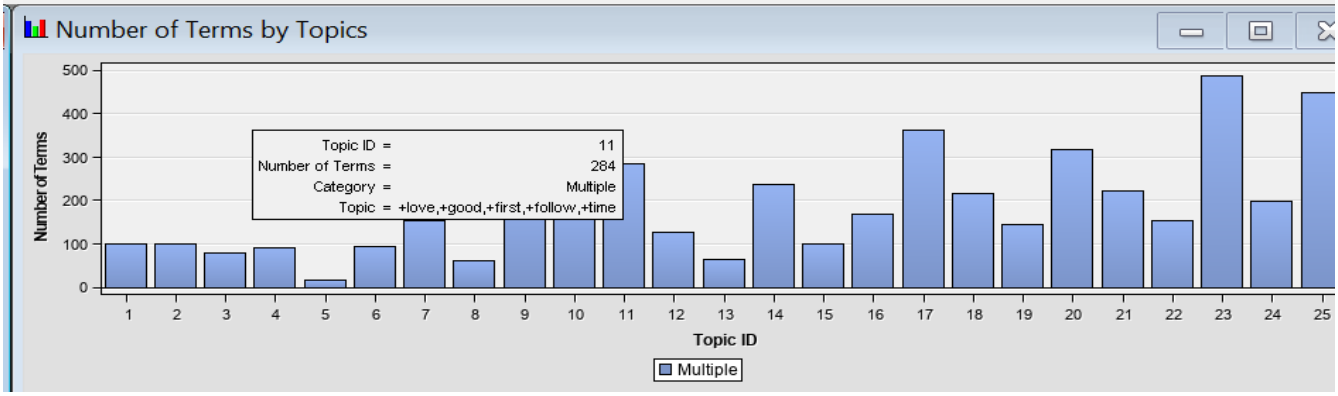
APPENDIX A: PROCESS FLOW

A.(i)



A.(ii)





APPENDIX B: MODEL RESULTS

TRAINING RESULTS

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Train: Average Profit for Label	Train: Total Profit for Label	Valid: Sum of Frequencies	Valid: Misclassification Rate	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Average Squared Error	Valid: Root Average Squared Error	Valid: Divisor for VASE
Y	Boost2	Boost2	Gradient ...	Label	Text Label	0.183619	6951	0.172349	0.982522	1713.118	0.123228	0.351039	13902	6951	5.955891	41399.4	2979	0.183619	0.958367	781.7577	0.131211	0.362231	5958
	Tree	Tree	Decision ...	Label	Text Label	0.193353	6951	0.173212	0.965986	1751.94	0.126021	0.354994	13902	6951	5.953206	41380.74	2979	0.193353	1	827.0744	0.138817	0.372582	5958
	Reg	Reg	Regressi...	Label	Text Label	0.201746	6951	0.193785	0.999506	1914.656	0.137725	0.371113	13902	6951	5.9556	41397.37	2979	0.201746	0.993395	836.7765	0.140446	0.374761	5958
	HPDMFo...	HPDMFo...	HP Forest	Label	Text Label	0.574018	.	0.502949	.	.	.	.	.	.	.	2125.54	.	0.574018	.	.	.	.	.

Event Classification Table

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Tree	Decision Tree	TRAIN	Label	Text Label	391	2663	813	3084
Tree	Decision Tree	VALIDATE	Label	Text Label	206	1119	370	1284
Boost2	Gradient Boosting	TRAIN	Label	Text Label	291	2569	907	3184
Boost2	Gradient Boosting	VALIDATE	Label	Text Label	150	1092	397	1340
Reg	Regression	TRAIN	Label	Text Label	219	2348	1128	3256
Reg	Regression	VALIDATE	Label	Text Label	119	1007	482	1371
HPDMForest	HP Forest	TRAIN	Label	Text Label	20	3250	226	3455
HPDMForest	HP Forest	VALIDATE	Label	Text Label	221	1135	354	1269

TEST RESULTS

Frequency| Percent | Row Pct | Col Pct	hateful	normal	Total
HATEFUL	949	621	1570
	16.97	11.10	28.07
	60.45	39.55	
	66.36	14.92	
NORMAL	481	3542	4023
	8.60	63.33	71.93
	11.96	88.04	
	33.64	85.08	
Total	1430	4163	5593
	25.57	74.43	100.00