



CS4242 Social Media Computing

Final project

# STOP THE CYBERBULLYING!

Classification and Identification of Toxic Tweets

Amelia Peh Yingqi | A0071186E

Wang Zhe | A0111785E

## Executive summary

Cyberbullying is growing problem in online communities, which has a large social cost due to the negative impact on the mental health of the victims. We develop a web-based application to classify toxic tweets to aid in identifying and investigating cyberbullying. The focus of the app is on user experience and ease of interpretation of the results, so as to make monitoring cyberbullying accessible to the masses. The user interface takes in keywords from users and runs the tweet miner to return relevant tweets, which are classified as toxic or non-toxic and are displayed on the interface. A list of words by term component strength based on topic modelling by LDA and LSI is also displayed in the tabs to allow users to use their human judgement to interpret the topics. The classifier uses a Naïve Bayes classifier algorithm and performs well on validation comment data, as well as real life tweet examples; although we note a slight over classification on tweets with profanities but are not derogatory in nature. 3 topics was determined to be suitable for the application for topic modelling by LDA and LSI; and both LDA and LSI give similar performance in terms of stability and the number of topics, although the topics may be different.

## 1. Introduction

### 1.1 Motivation

The increasing popularity of social media also leads to more prevalent cyberbullying as a side effect. Cyberbullying can negatively impact the victim, many of which are children and teenagers, and can lead to depression, anxiety, low self-esteem, or even attempted suicide. (Kaltiala-Heino et al. 2000) Cyberbullying when gone unchecked leads to a huge social cost to the society and thus making monitoring and investigation of cyberbullying an important task.

### 1.2 Problem Statement

Cyberbullying is growing problem in online communities, and the detection of cyberbullying is a challenging task left mostly to social media platform providers which may not prioritise monitoring cyberbullying over other priorities like profits and freedom of speech.

In this application, we bring the ability to check up on toxic tweets in a certain community or locality to the wider group of people, such as parents, schools and peers. In so doing, we bring the concept of crowd sourcing to monitor cyberbullying. This helps to increase transparency on cyberbullying prevalence, expose bullying patterns, and allow help to be rendered to victims.

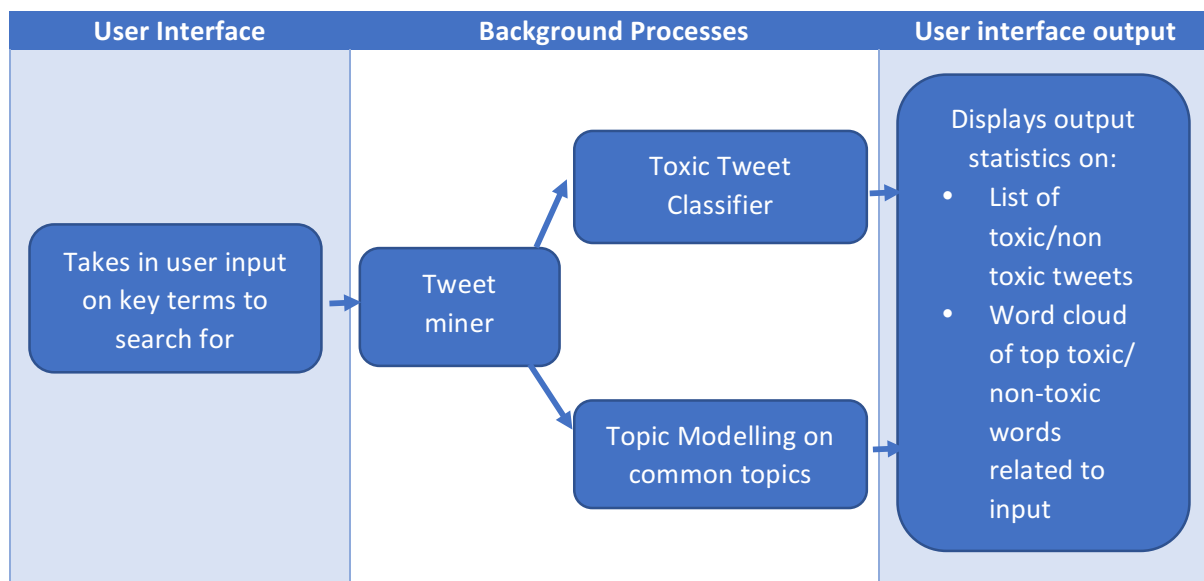
### 1.3 Current State of Art

The use of machine learning in harassment detection from comments and chat datasets has been widely studied by researchers. Yin et al. (2009) used a bag-of-words approach examined a baseline text mining system and improved by including sentiment and contextual features, while Nahar et al. (2013) made use of social network features to improve classifier accuracy. More recently, deep learning is employed on tweets to detect hate speeches (Badjatiya et al., 2017). However, these researches are not made accessible to the masses in the form of an application, or other user-friendly form and thus, remains in the realm of academia.

## 2. Methodology

### 2.1 Program Structure

The Application consists of the following components:



## 2.2 Training Data Set

The training data is a set of Wikipedia comments human labelled for toxicity, severe toxicity, obscene, insult, threat, identity hate. This forms the basis for building the toxic tweet classifier.

## 2.3 Toxic Comments Classifier

### 2.3.1 Data Pre-processing

The toxic comment classifier first pre-processes the input data for the following:

- Convert to lowercase
- Remove URLs
- Remove IP addresses
- Remove Mentions
- Remove hashtag symbol
- Remove time
- Remove punctuations
- Word tokenize and stemming
- Remove stop words
- Remove low frequency words

Feature selection was performed using document frequency thresholding and was completed as part of the pre-processing step. Document frequency (DF) is the number of documents containing the term. As terms with high DF are preferred during Text classification, terms with DF less than 3 was removed from the corpus.

### 2.3.2 Model Building and Optimisation

The model optimisation was performed based on the following procedure:

- Pre-processed data are split into 80% training and 20% validation set.
- Training data is used for model selection and optimisation based on 10-fold cross validation.
- Optimal model from cross validation F1 score is trained using the entire 80% training data set.

- d. The 20% validation data set is used to get the final accuracy of the classifier.

The final trained model is called in the web-based application to classify fresh tweets mined from Twitter.

## 2.4 Topic Modelling

Two different methods of topic modelling were explored:

- Latent Dirichlet Allocation (LDA)  
LDA is a probabilistic model that suggests a set of global topics (i.e., a set of discrete distributions over words) and a set of document topics (i.e., a set of discrete distributions over topics, one distribution per document). The output, a set of words each for the number of topic specified, helps to identify possible themes and topics.
- Latent Semantic Indexing (LSI), also known as Latent Semantic Analysis (LSA).  
LSI is based on linear algebra and makes use of Singular Value Decomposition (SVD) on the Document-Term Matrix to group words into topics.

The results from the 2 different methods were examined and compared in Section 3.2.

## 2.5 Tweet Miner

The tweet miner runs on Python and it calls the Twitter API to retrieve tweets related to the keywords that the user inputs.

## 2.6 Web-based Application

The web-based application combines all the modules, the Tweet miner, the Toxic Comment Classifier, and Topic Modeller into a user interface. The sequence of how the web application works is as follows:

- a. User interface takes in keywords from users
- b. Tweet miner is run to extract relevant tweets from the Twitter API based on keywords
- c. Tweets are classified as toxic or non-toxic
- d. Topic modelling by LDA and LSI is performed on the retrieved tweets
- e. User interface displays the tweets returned, classification on toxic or non-toxic, and top 10 words for each topic by LDA and LSI

# 3. Results and Discussion

## 3.1 Toxic Comments Classifier

The text classifier built on labelled toxic comments are optimised using various machine learning models using 10-fold cross validation using the training data. Based on models trained on the individual classifiers, a fusion model based on majority voting is built and tested using a holdout validation set. The final fusion model was able to achieve a precision, recall and F1 of 0.8993, 0.8972, 0.8983 respectively.

	Precision	Recall	F1 Score
Naïve Bayes Classifier	0.8826	0.8826	0.8826
Random Forest Classifier	0.8272	0.8195	0.8233
K Neighbours Classifier	0.7353	0.7301	0.7327
Support Vector Classifier	0.8860	0.8858	0.8859
XGBoost	0.8468	0.8332	0.8399
Fusion Model	0.8993	0.8972	0.8983

Table 1: Classification accuracy of various machine learning models on training data set

The classifier performs well for the comments in terms of accuracy. As there are no labelled tweets for the validation of the classifier performance, we manually inspect the outcome of the classified tweets. Based on human judgement, the classifications are largely accurate on sample tweets mined and input into the classifier. A few examples using the keyword “Trump”, for American President Donald Trump, are listed in table 2 as illustration.

Tweet content	Classification
“Trump is a fat orange asshole and Pence is a closet gay hiding behind his mothers//wife’s skirt!!!”	Toxic
“@realDonaldTrump Hey motherfucker! You repugs are gonna find at WAR at home! You are not president and never will be”	Toxic
“@nytimes Trump does what he wants and earned the right to do this upon winning the last election Trump supporters d...”	Non-toxic
“I just signed this petition telling Congress not to pass a dangerous proposed AUMF that would give President Trump...”	Non-toxic

Table 2: Inspection of tweet content and classifier results

However, we note a proportion of false positives, i.e. classification of tweets classified as toxic because of the profanities used but are not derogatory or cyberbullying targeted at others inherently. These tweets are typically just cursing their luck or contain profanities just a manner of speech to emphasize some points. This is highly dependent on the human label used to build the training model; i.e. at which threshold is a comment considered toxic or just a comment with profanities. Nevertheless, we believe that the cost of over labelling (having more toxic comments to sort through) is still lower than that of under labelling (having cyberbullying go undetected).

## 3.2 Topic Modelling

### 3.2.1 Selecting number of topics

To test the LDA and LSI models created, a random sample of 200 comments from the training data set for the toxic tweet classifier was used as input for topic modelling. Typically, the number of topics can be selected based on the perplexity of the LDA model, where perplexity is a statistical measure of how well a probability model predicts a sample. However, a study by Chang et al. (2009), shows that perplexity and human judgment are often not correlated and thus the model with the lowest perplexity may not be the most semantically meaningful. Hence, the number of topics was chosen by rule of thumb and intuition instead of perplexity.

Number of topics were set to 2 and 3 for comparison purposes and the results of terms by term component strength in the topic are plotted in Figure 1a and b and 2a to c. From the results, 2 topics does not give a good distribution of topics as the terms for both topics (figure 1a and 1b) are somewhat similar. 3 topics in contrast have more distinct related words in each of the topics, where topic 1 is largely general disagreement, topic 2 is polite and helpful comments, while topic 3 is negative war-related comments.

3 topics will be displayed in the web application as we believe that 2 topics is too few for good differentiation of topics, while 4 or more may be too much due to the limit of 100 tweets to be mined per minute from the Twitter API which may lead to sparsity and repeated topics.

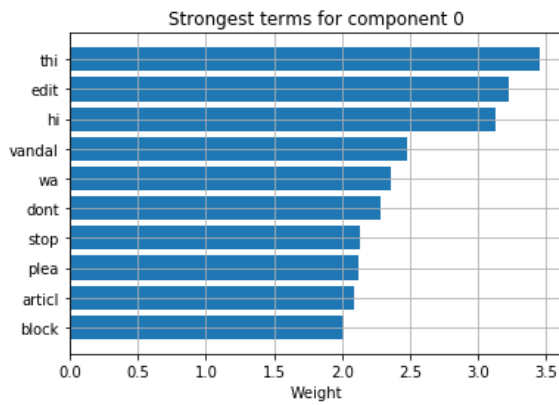


Figure 1a: Topic 1 out of 2 LDA results

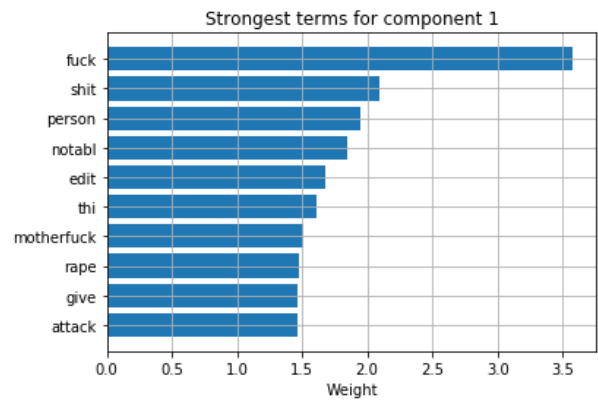


Figure 1a: Topic 2 out of 2 LDA results

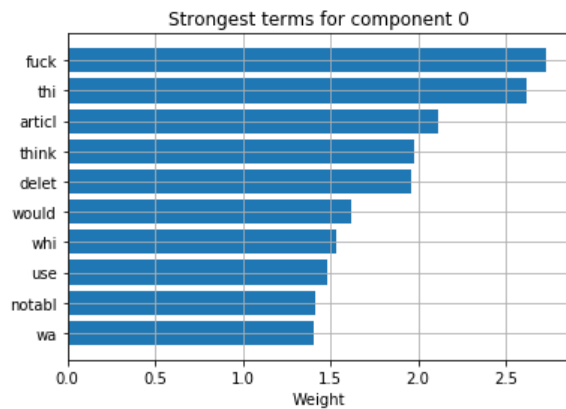


Figure 2a: Topic 1 out of 3 LDA results

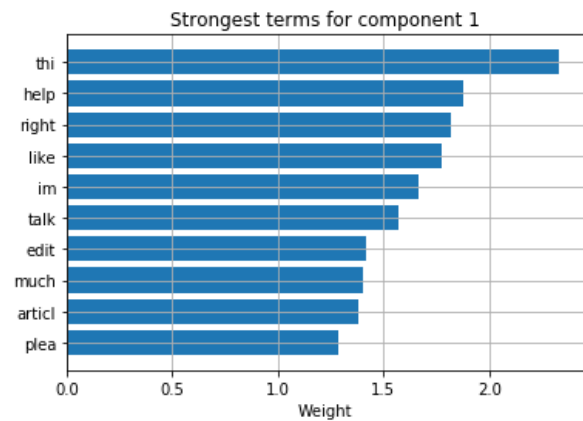


Figure 2b: Topic 2 out of 3 LDA results

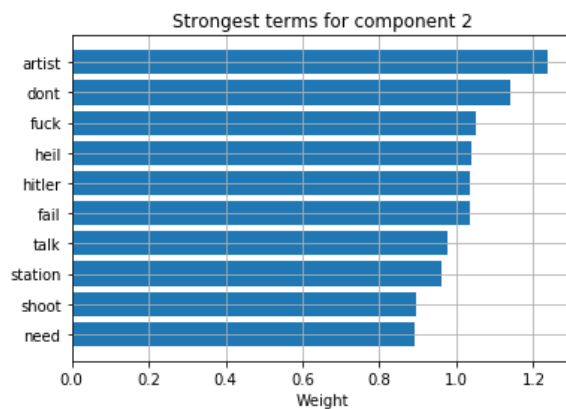


Figure 2c: Topic 3 out of 3 LDA results

### 3.2.2 Comparison of LDA vs LSI

Between the two different methods of topic modelling, LDA and LSI, the performance is not significantly different in terms of stability and the number of topics for the best analysis. Although LDA is a probabilistic model, by setting the random state and using the same input data, we can get topics that are similar between each run of the LDA. There are no such issues with LSI as the method is deterministic in nature.

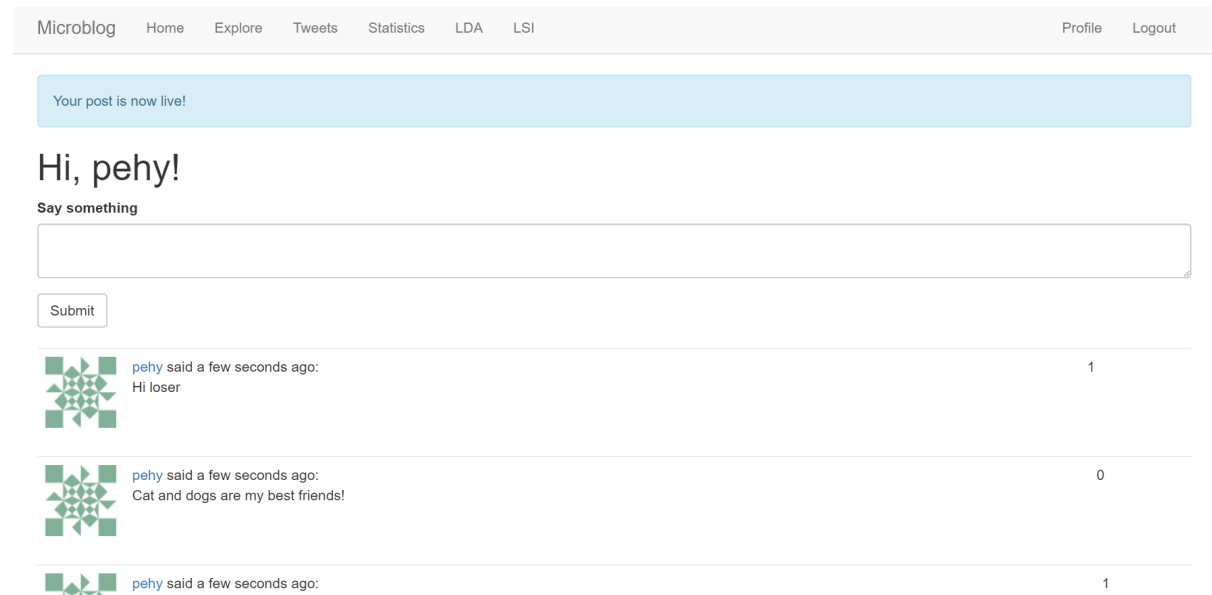
The topics resulting from the LDA and LSI however can be different with the same input data. As such, we include both LDA and LSI into the web application for users to interpret using human judgement.

### 3.3 Web-based Application

The focus of the web-based application is user experience and ease of interpretation of the results. We list below snapshots of the main functionalities of the application:

#### 3.3.1 Home Page

This page allows users to input microblog contents and it displays the content with the classification on toxicity.

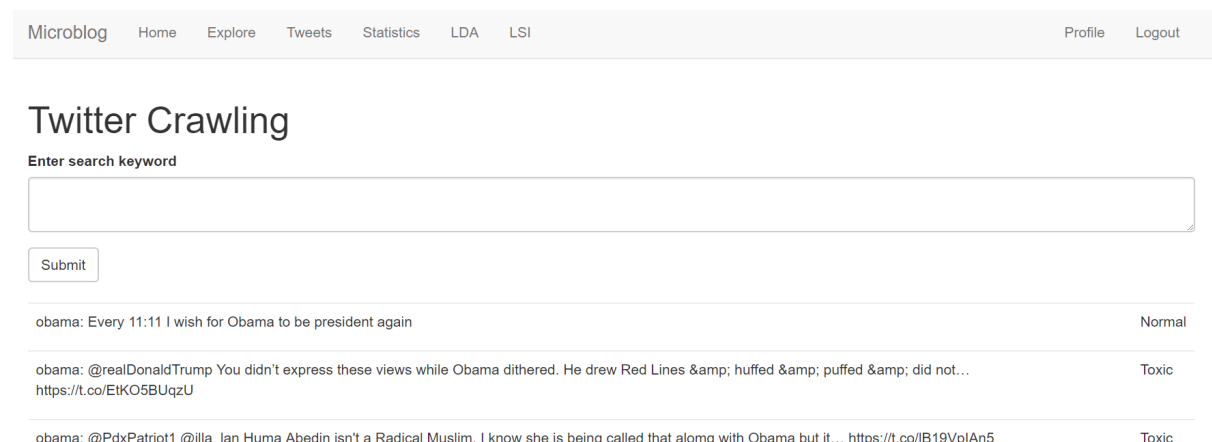


Profile Picture	Username	Time	Text	Toxicity Score
	pehy	said a few seconds ago:	Hi loser	1
	pehy	said a few seconds ago:	Cat and dogs are my best friends!	0
	pehy	said a few seconds ago:		1

Figure 3a: Twitter Crawling Page which allows user to input keywords

#### 3.3.2 Twitter Crawling Page

This page allows users to input keywords and the page returns with relevant tweets listed with the tweet content and the classification on toxicity.



Username	Text	Toxicity Score
obama:	Every 11:11 I wish for Obama to be president again	Normal
obama: @realDonaldTrump	You didn't express these views while Obama dithered. He drew Red Lines & huffed & puffed & did not... https://t.co/EtKO5BUqzU	Toxic
obama: @PdxPatriot1 @illa_lan	Huma Abedin isn't a Radical Muslim. I know she is being called that along with Obama but it... https://t.co/IB19VplAn5	Toxic

Figure 3b: Twitter Crawling Page which allows user to input keywords

#### 3.3.3 Statistics Page

This page displays the results of the word count and TF-IDF of the words by popularity.

## Tweets Statistics



Figure 3c: Web interface for LDA results

### 3.3.4 Latent Dirichlet Allocation and Latent Semantic Indexing Page

This page displays the results of the LDA and LSI for users to interpret. LSI page is not shown in the diagram as it is similar to the LDA page.

## Latent Dirichlet Allocation



Figure 3d: Web interface for LDA results

## 4. Future work

We note the limitation of the training data set used for the classifier training, which is labelled based on Wikipedia comments, rather than real tweets, due to a lack of ground truth labelled for cyberbullying for tweets available online. A tweets data set labelled for toxicity will definitely help in improving classifier accuracy and allows the model to be validated with tweets.



## 5. Conclusion

In conclusion, we develop a web-based application to classify toxic tweets to aid in identifying and investigating cyberbullying. The focus of the app is on user experience and ease of interpretation of the results, so as to make monitoring cyberbullying accessible to the masses. The user interface takes in keywords from users and runs the tweet miner to return relevant tweets, which are classified as toxic or non-toxic and are displayed on the interface. A list of words by term component strength based on topic modelling by LDA and LSI is also displayed in the tabs to allow users to use their human judgement to interpret the topics. The classifier performs well on validation comment data, as well as real life tweet examples, although we note a slight over classification on tweets with profanities but are not derogatory in nature. 3 topics was determined to be suitable for the application for topic modelling by LDA and LSI; and both LDA and LSI give similar performance in terms of stability and the number of topics, although the topics may be different.

## References

- Badjatiya, P., Gupta, S., Gupta, M. & Varma, V. (2017) Deep Learning for Hate Speech Detection in Tweets, Proceedings of the 26th International Conference on World Wide Web Companion, April 03-07, 2017, Perth, Australia
- Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. M. (2009) Reading Tea Leaves: How Humans Interpret Topic Models. Neural Information Processing Systems.
- Kaltiala-Heino, R., Rimpel, M., Rantanen, P. & Rimpel, A. (2000), 'Bullying at school-an indicator of adolescents at risk for mental disorders', Journal of Adolescence 23, 661–674.
- Nahar, V., Li, X. & Pang, C. (2013), An effective approach for cyberbullying detection, Journal of Communications in Information Science and Management Engineering (CISME) 3, 238–247.
- Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L., "Detection of Harassment on Web 2.0," in Proc. Content Analysis of Web 2.0 Workshop, Madrid, Spain, 2009.