Cyber bullying detection on social media

Jonathan Kow Yee Seng, Siti Zulaikha Dollah, Syed Rusyaidi Syed Putra,Huang Huang
University of Malaya, Kuala Lumpur, Malaysia

Abstract—Cyberbullying is an act of bullying with the use of technology as a medium. The exponential growth of social media has exposed many people especially young people to the risk of being target of cyberbullying. With the help of machine learning, we can detect language patterns from the cyberbullying posts and develop a model to automatically detect cyberbullying content.

This project uses data from Formspring.me, a question-and-answer formatted website. The data is labelled by Amazon's Mechanical Turk, a web service. After data preparation and pre-processing, Scikit-learn, a python library is used to train a model to classify if cyberbullying is present in the post by recognising bullying content based on its insult words as features. Four machine learning techniques namely logistic regression, decision tree, random forest and support vector machine are used to train the model. As a result, all four algorithms achieve more than 78% of accuracy in identifying the true positive where support vector machine achieves the best performance with a score of 87% of accuracy in identifying the true positive. Furthermore, the context of cyberbullying post is investigated and categorised into five categories namely swearing, abusive, sexism, vulgar, and racism. Last but not least, a study on the severity level and the context of cyberbullying content analysed swearing as the most frequent category of cyberbullying in this dataset.

1 Introduction

1.1 Background

Cyberbullying is the act of bullying which conducts on digital devices like smart phone, computers, tablets and so on where internet is accessible. Common means of cyberbullying can take place are text, forum discussion and the mainstream social media with the condition that people are able to view, participate, or share their opinion. Given the rise of industry revolution, social media is getting more mainstream than ever, social media users are growing exponentially including people from all walks of life. Sending, posting, or sharing potentially harmful content, false information or, leaking private information of others that would cause embarrassment or humiliation are examples of cyberbullying that happens in our daily life especially in the abstract world of social media. Although law against cyberbullying is established, but criminals are way too many to tackle or being brought to justice to face charges. Moreover, this phenomenon leaves people especially young people exposed to the risk of being the target of cyberbullying. Tragically, in more extreme circumstances where young victims are vulnerable and under protection of no guardians, they ended up commit suicide as a way of escaping from the haunt of cyberbullying.

1.2 Problem Statement

The number of social media company has risen but approaches to prevent cyberbullying have not been well-equipped by the application, leaving cyberbullies to roam across internet and attacking innocent people with no justice. With this being said, prevention measures and law protection are not keeping up with the criminal records of cyberbullying.

1.3 Objective

In fact, there are numerous remarkable researches conducted to overcome cyberbullying specifically on detection of cyberbullying with the help of machine learning to study the sentiment and contextual features from the conversation or medium. This project on the other hand will tackle this problem by performing classification on social media data to detect the act of cyberbullying. In particular, the objectives of this project are:

1) To classify posting susceptibility to cyberbullying with certain degree of confidence.
2) To investigate the context of discrimination on posting susceptible to cyberbullying.
3) To study the severity of cyberbullying based on the context of cyberbullying.

2 Data collection

This section will describe the details of data collection and data labelling in this project.

2.1 Data Origin

The origin of data in this project is a question and answer-based website, Formspring.me where users openly invite others to ask and answer questions. The feature of anonymity option which allow users to post question anonymously to other user's profile. This dataset is provided in Kaggle [1] which represented 50 ids from Formspring.me that were crawled in Summer 2010. The columns are user id, post, question of post, answer of post, and response of labelling cyberbullying, bully words, and the severity level in a range of 10 from three different respondents.

2.2 Data Labeling

The labelling cyberbullying work was determined by Amazon's Mechanical Turk service, an online marketplace that allows requestors to post tasks which are then completed by paid workers. The labelling work was done according to the following question:

1) Does this post contain cyberbullying (Yes, 1 or No, 0)?
2) On a scale of 1 (Mild) to 10 (Severe) how bad is the cyberbullying in this post (0 for no cyberbullying)?
3) What words or phrases in the post(s) are indicative of the cyberbullying (NaN for no cyberbullying)?

3 Data Preparation

This section will describes data preparation process before the stage of modelling.

3.1 Data Cleaning

The first cleaning process involves dropping any missing data from post, questions of post, or answer of post. Then, the convert of 'Yes' and 'No' value to integer '1' and '0' for conveniences of further processing. Next, the dataset is filtered so that the response of '0' from all respondents agreed upon are accepted and at least two response of '1' from respondents are accepted to be '1'. This leads to the creation of a new column 'overall_ans' which conclude the filter step above aggregating three responses into one overall decision. In addition to that, severity from three respondents are averaged to a new column named 'overall_severity' for aggregating into one decision value.

3.2 Data Preprocessing

Preprocessing process takes places with four common text preprocessing approaches. Firstly, tokenizing the data. This is a process where strings are tokenized or split into a list of tokens or small parts in order for the machine to be able to understand and capture the pattern. In this step, text are split based on whitespace and punctuation with the help of a machine learning module, punkt which is a pre-trained model that tokenize words and sentences. Next, the text is normalized to convert to its canonical form. This process helps to group words which presented in different forms but same meaning. One of the famous techniques of normalization, lemmatization is used to analyses the structure of the word and its context to convert to a normalized form with the help of module that make use of a lexical database for English language that helps the script determine the base word, wordnet [2]. Last but not least, the noise is removed from the data with the help of regular expression and a stopwords resource from python module. Example of noise includes hyperlink, numbers, stop words and so on.

After preprocessing step, the data comprises 11609 values where only 776 values (nearly 7% of data) are positive values (cyberbullying, '1'). A severe imbalanced property of data is observed, and extra care is essentially needed for modelling and analysing the reading of result.

4 Model Preparation

This section describes the preparation for preprocessed text to be ready to split into training, validating, and testing purpose.

4.1 Resampling

Due to the severe imbalanced property of the data, this project introduces resampling technique which will be applied to training set to minimize this potential disadvantage of data. Resampling is a technique that consists of drawing repeated samples from the original data samples. The resampling approaches conducted are oversampling the minority class, a type of data augmentation for the minority class and referred to as Synthetic Minority Oversampling technique (SMOTE) described by Nitesh Chawla [3]. Furthermore, this paper on SMOTE also suggested combining SMOTE with random undersampling of the majority class which this project takes as one of the approaches. With the help of a python library 'imbalanced-learn', the minority class (positive value, '1' as cyberbullying) is oversampling to have 10 percent of the number of examples of the majority class whereas the majority class (negative value, '0', as no cyberbullying) is undersampling to reduce the number of examples in the majority class to have 50 percent more than the minority class.

4.2 Feature Extraction

This project illustrates a bag-of-words model in which a way of extracting features from text for use in modelling where bag-of-words is a representation of text that describes the occurrence of words within a document [4]. In this step, a collection of text posts is converted to matrix of token counts and a creation of words vocabulary. Then, due to the reason that large counts of certain words might not be meaningful in the encoded vectors. Thus, Word frequencies are calculated by one of the popular approaches, Term Frequency – Inverse Document (TF-IDF) which summarizes how often a given word appears within a document and downscales words that appear a lot across documents in order to capture meaningful pattern [5]. This process is conducted with the help of python library, scikit-learn

5 Model Evaluation

This section discusses the machine learning techniques used, and the analysing of statistical results indicating the learning experiment. In addition, Python language will be used to evaluate the model. The machine learning techniques or algorithms conducted are logistic regression, decision tree, random forest and support vector machine. Also, model evaluation is based on three set of data: training set (70%), validation set (20%), and testing set (10%).

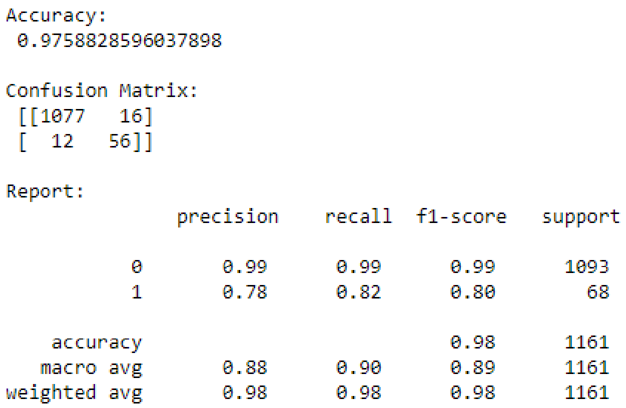
In addition, due to the reason that the dataset is imbalanced, the model evaluation will focus heavily on the reading of the accuracy of positive value, and recall of positive value, true positive, and false negative as the overall accuracy is initially high even if prediction of all value is set to 0 or negative. As shown in code file, only with prediction of all negative value (as no cyberbullying), the model will still achieve a high overall accuracy of 94.14%. Usual way of interpretation will be discarded and mainly focus on how well the algorithm capture the positive value (true positive) and how bad the model missed out to capture the positive value (false negative).

5.1 Logistic Regression

Logistic regression is a supervised learning classification algorithm that can be used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there would only be two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (success/yes) or 0 (failure/no). To achieve the main objective of this project, Logistic Regression is used as one of the models to predict the if cyberbullying is present in social media post. Logistic regression is refined with a threshold of 0.4 probability in a way of saying, a post has 0.4 probability of classifying as cyberbullying (low threshold). This is due to the reason that, classifying a cyberbullying post as negative (as no cyberbullying) is more costly that classifying a non-cyberbullying post as positive (as cyberbullying). Intuitively, the algorithm will be less possibly to miss out positive value (as cyberbullying).

The result from testing set shows that the model achieves a high recall of 82%, F1-score of 80%, and precision of 78% for positive value ('1', as cyberbullying); and an expected high reading of precision of 99%, recall of 99%, and F1-score of 99% for negative value ('0', as no cyberbullying). Also, for reporting purpose, the overall accuracy is 97.59%. Summary of report for logistic regression is shown in figure 1.

In conclusion, logistic regression achieves a good performance in detecting cyberbullying.



```
Accuracy:
 0.9758828596037898

Confusion Matrix:
 [[1077   16]
 [  12   56]]

Report:
               precision    recall  f1-score   support

           0       0.99      0.99      0.99      1093
           1       0.78      0.82      0.80        68

    accuracy                           0.98      1161
   macro avg       0.88      0.90      0.89      1161
weighted avg       0.98      0.98      0.98      1161
```

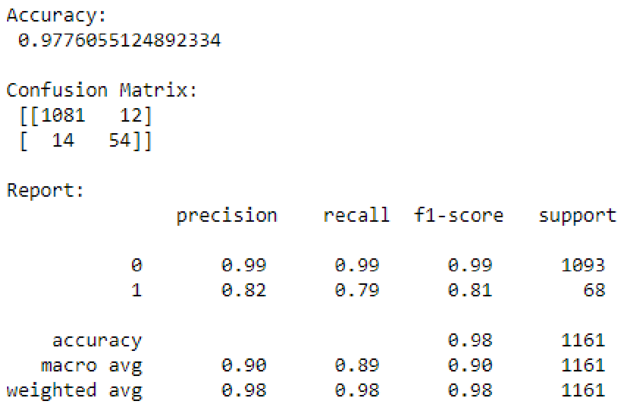
Fig. 1: Summary of Logistic Regression

5.2 Decision Tree Classifier

Decision tree is a decision support tool that uses tree-like graph includes the possibilities of event outcomes, information gain, and resource costs. It is one of the algorithms that contains conditional control statements. Tree based method empower predictive models with advantages of high accuracy and easily interpreted. Therefore, decision tree classifier is one of the approaches to classify if cyberbullying is present in a post based on its conditional properties.

The result from testing set shows that the model achieves a relatively high recall of 79%, F1-score of 81%, and precision of 82% for positive value ('1', as cyberbullying); and an expected high reading of precision of 99%, recall of 99%, and F1-score of 99% for negative value ('0', as no cyberbullying), and an overall accuracy of 97.76%. Summary of report for decision tree classifier is shown in figure 2.

In conclusion, decision tree classifier achieves a relatively good performance in detecting cyberbullying



```
Accuracy:
 0.9776055124892334

Confusion Matrix:
 [[1081   12]
 [  14   54]]

Report:
               precision    recall  f1-score   support

           0       0.99      0.99      0.99      1093
           1       0.82      0.79      0.81        68

    accuracy                           0.98      1161
   macro avg       0.90      0.89      0.90      1161
weighted avg       0.98      0.98      0.98      1161
```

Fig. 2: Summary of Decision Tree

5.3 Random forest Classifier

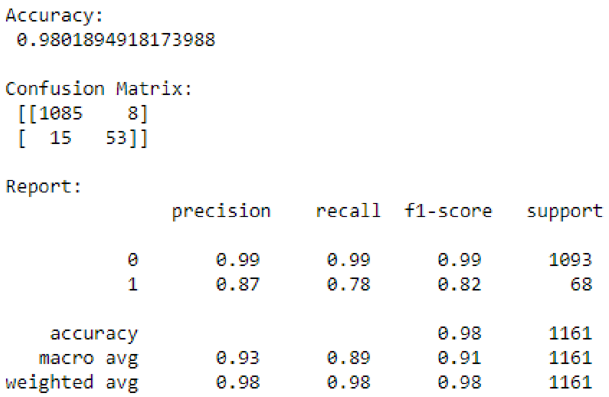
Random Forest Classifier are used to train our model to classify if post with cyberbullying. Random Forest is one of the most known and effective machine learning algorithms in data mining and particularly in text classification. A difference between Random Forest and Decision Tree is decision tree is built on an entire dataset using all features whereas Random Forest randomly selects specific features to build multiple decision trees and get the average result. Each of the decision tree consist class prediction and the most votes of the prediction will be selected as algorithm model. An addition feature is it uses bootstrap sampling to extract a number of training samples and group the features from the original training set, establishes a plurality of unpruned decision trees. Then, combines the decision trees to form a random forest model. As a result, the randomness increases the diversity of decision trees and makes the resulting integrated model have better classification performance. Due to the feature of random forest, it is considered as one of the approaches to capture the pattern of post with cyberbullying.

The result from testing set shows that the model achieves a relatively high recall of 78%, F1-score of 82%, and precision of 87% for positive value ('1', as cyberbullying); and an expected high reading of precision of 99%, recall of 99%, and F1-score of 99% for negative value ('0', as no cyberbullying), and an overall accuracy of 98.02%. Summary of report of Random Forest classifier is shown in figure 3.

To sum up, Random Forest Classifier is pretty accurate in detecting cyberbullying.

5.4 Support Vector Machine

Support vector machine (SVM) is a technique that can be used for both classification and regression problem. To achieve the main objective, SVM is used to perform classification by finding the hyper-plane that differentiates the two classes. The motivation to consider SVM as one of



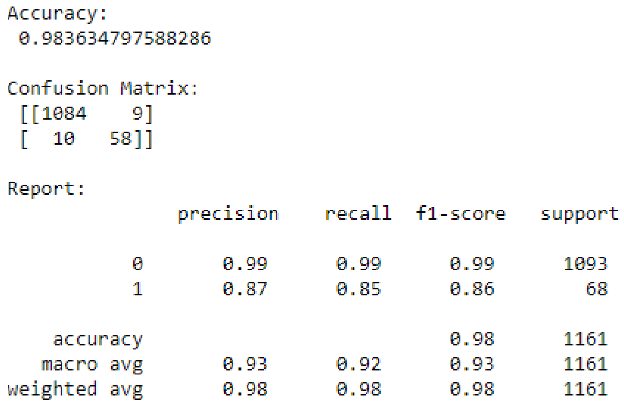
```
Accuracy:
 0.9801894918173988

Confusion Matrix:
 [[1085    8]
 [  15   53]]

Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1093
           1       0.87      0.78      0.82        68

    accuracy                           0.98      1161
   macro avg       0.93      0.89      0.91      1161
weighted avg       0.98      0.98      0.98      1161
```

Fig. 3: Summary of Random Forest

the approaches is due to its property of being an optimal margin classifier. With this being said, this algorithm will try to find a decision boundary that maximizes the geometric margin and results in a very confident set of predictions on modelling. In particular, SVM will measure the maximum distance between post with cyberbullying and post without cyberbullying in order to form the optimal decision boundary line. Therefore, linear SVM is chosen for modelling to classify if cyberbullying is present in post.

The result from testing set shows that the model achieves a relatively high recall of 85%, F1-score of 86%, and precision of 87% for positive value ('1', as cyberbullying); and an expected high reading of precision of 99%, recall of 99%, and F1-score of 99% for negative value ('0', as no cyberbullying), and an overall accuracy of 98.36%. Summary of report for SVM is shown in figure 4.

All in all, to no surprise, SVM achieves a great performance in detecting cyberbullying and it trained a very reliable model.



```
Accuracy:
 0.983634797588286

Confusion Matrix:
 [[1084    9]
 [  10   58]]

Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1093
           1       0.87      0.85      0.86        68

    accuracy                           0.98      1161
   macro avg       0.93      0.92      0.93      1161
weighted avg       0.98      0.98      0.98      1161
```

Fig. 4: Summary of SVM

5.5 Summary

Last but not least, a conclusion to model evaluation is SVM achieves the best performance overall on this dataset with the highest accuracy for positive value of 87% which considered as one of the main performance metrics as this metric captures how good the post with cyberbullying is captured, and the highest recall for positive value of 85%. This means it correctly classifies the most post with cyberbullying and post with no cyberbullying.

6 Dissection: Cyberbullying

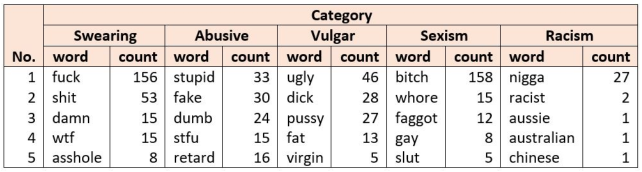
This section will discuss the result of investigation of the context of discrimination in cyberbullying post (second objective) and the level of severity of cyberbullying content based on its context (third objective).

6.1 Context of Discrimination

In this step, the feature of cyberbullying is extracted based on the potential bully words in cyberbullying post. Furthermore, the bully words are categorised into five categories based on the bully words. The categories are swearing, abusive, vulgar, sexism and racism.

1) Swearing – An expression of strong feeling towards something, generally considered to be language that is strongly impolite, rude, or offensive.
2) Abusive – The use of remarks intended to be demeaning, humiliating, mocking, insulting, or belittling people.
3) Vulgar – A practice that make explicit and offensive reference to sex or bodily functions.
4) Sexism – Prejudice, stereotyping or discrimination on the basis of sex typically against woman.
5) Racism – Prejudice, discrimination, or antagonism directed against a person or people on the basis of their membership of a particular racial or ethnic group.

Example of bully words are categorised into a table for better alignment as shown in figure 5.



	Category									
	Swearing		Abusive		Vulgar		Sexism		Racism	
No.	word	count	word	count	word	count	word	count	word	count
1	fuck	156	stupid	33	ugly	46	bitch	158	nigga	27
2	shit	53	fake	30	dick	28	whore	15	racist	2
3	damn	15	dumb	24	pussy	27	faggot	12	aussie	1
4	wtf	15	stfu	15	fat	13	gay	8	australian	1
5	asshole	8	retard	16	virgin	5	slut	5	chinese	1

Fig. 5: Categories of bully words

In addition to that, wordcloud visualization technique is used to better observe the result as shown in figure 6.

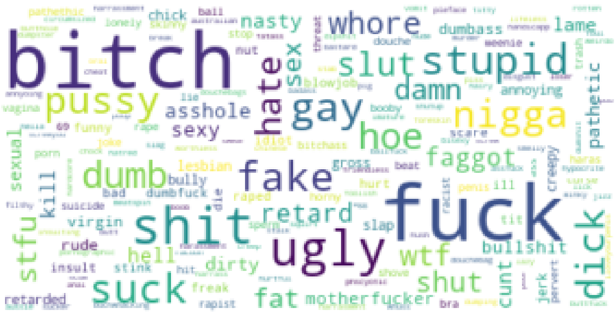


Fig. 6: Wordcloud of bully words

Overall, swearing consists of 34.6%, sexism consists of 27.7%, abusive 16.5%, vulgar 16.7% and racism 4.5% out of all cyberbullying post. It appears that swearing is most common practice in cyberbullying in this dataset. A pie chart is used for better visualization as shown in figure 7.

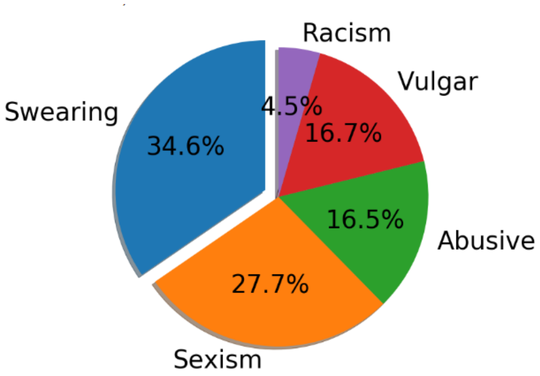
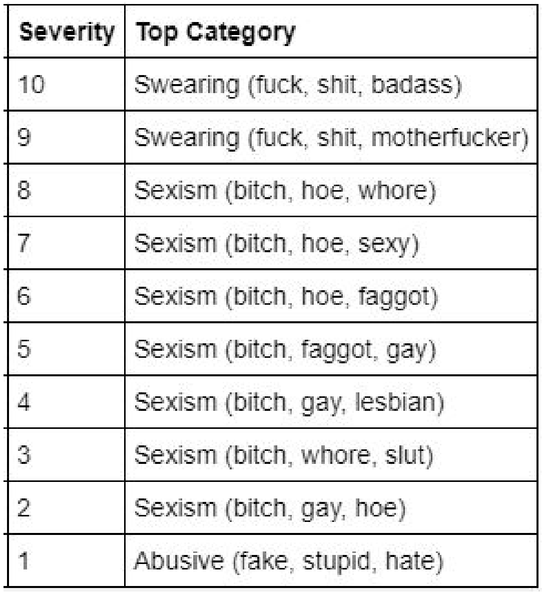


Fig. 7: Piechart for categories of bully words

6.2 Severity of Cyberbullying

In addition, this step will examine the severity level of cyberbullying based on the context of cyberbullying. Investigation is conducted to examine the most frequent category of bully words based on its severity level. As mention above, there are 10 severity level where ranging from 1 (mild) to 10 (severe). For better visualization, the result is tabulated as shown in figure 8.



Severity	Top Category
10	Swearing (fuck, shit, badass)
9	Swearing (fuck, shit, motherfucker)
8	Sexism (bitch, hoe, whore)
7	Sexism (bitch, hoe, sexy)
6	Sexism (bitch, hoe, faggot)
5	Sexism (bitch, faggot, gay)
4	Sexism (bitch, gay, lesbian)
3	Sexism (bitch, whore, slut)
2	Sexism (bitch, gay, hoe)
1	Abusive (fake, stupid, hate)

Fig. 8: Cyberbullying category for severity level

According to the result, swearing is most widely used in the high severity, then followed by sexism and lastly abusive in low severity in this dataset based on their severity level.

In conclusion, swearing is the most common practice in cyberbullying and ranks as the highest severity level according to the dataset.

7 Conclusion

In a nutshell, this project uses language-based method of detecting cyberbullying by classification. By capturing the pattern of cyberbullying based on the insult words as features, we are able to correctly identify 87% of posts that contain cyberbullying based on the best result from SVM algorithm. All our results indicate that the model is doing a reasonable job in detecting cyberbullying in Formspring posts.

Besides that, for future work, social media company has to develop or adopt cyberbullying detection in application so that the risk of users exposed to cyberbullying can be minimized

References

[1] S. Agrawal, "Formspring data for cyberbullying detection," https://www.kaggle.com/swetaagrawal/formspring-data-for-cyberbullying-detection, 26 January 2017.

[2] WordNet, "Wordnet interface," http://www.nltk.org/howto/wordnet.html.

[3] N. V. C. et al, "Smote: Synthetic minority over-sampling technique," https://arxiv.org/pdf/1106.1813.pdf, June 2002.

[4] Y. Goldberg, "Neural network methods in natural language processing," Synthesis Lectures on Human Language Technologies, April 2017.

[5] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=A564D6FFDFED929E546749DE73E9D30C?doi=10.1.1.115.8343&rep=rep1&type=pdf.