IMPLEMENTATION OF HYPERPARAMETER OPTIMISATION AND OVER-SAMPLING IN DETECTING CYBERBULLYING USING MACHINE LEARNING APPROACH

Wan Noor Hamiza Wan Ali[1], Masnizah Mohd[2], Fariza Fauzi[3], Kiyoaki Shirai[4], Muhammad Junaidi Mahamad Noor[5]*

[1,2,3]Center for Cyber Security, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia,
Bangi, Selangor Malaysia

[4]School of Advanced Information Science, Japan Advanced Institute of Science and Technology,
Nomi, Ishikawa Japan

[5]INVOKE Solutions Sdn Bhd, Sungai Besi, Kuala Lumpur, Malaysia

Email: p93244@siswa.ukm.edu.my[1], masnizah.mohd@ukm.edu.my[2*](corresponding author),
fariza.fauzi@ukm.edu.my[3], kshirai@jaist.ac.jp[4], junaidi@invokeisdata.com[5]

DOI: https://doi.org/10.22452/mjcs.sp2021no2.6

ABSTRACT

Online social networks have become a necessity to everyone around the world. Particularly, online social networks have enabled us to connect to one another regardless of time, for as long as we have social media and social networking as platforms for broadcasting information and communicating, respectively. However, this evolution has resulted in people possibly committing various cybercrimes, such as cyberbullying. To address this issue, machine learning can be utilised to counter cyberbullying in online social networks. Thus, this study proposed a framework with a set of features consisting of word and character term frequency–inverse document frequency and word embedding by using Word2vec and six types of list terms: profane words, proper nouns, negation words, 'allness' term, diminisher words and intensifier words. These features were divided into four groups before being fed into the linear support vector classifier to train our model using ASKfm as data set in hyperparameter tuning and over-sampling environment. Results indicated that the proposed framework provided significant outcomes, in which the highest percentage of area under curve is 99.24% and F-measure is 97.38% as performed by our trained model.

Keywords: Classification, Cyberbullying, Feature Extraction, Hyperparameter Optimisation, Machine Learning, SMOTE, TF-IDF, Word Embedding

1.0 INTRODUCTION

The Internet and social media have evolved, grown rapidly, and become one of the necessities in daily life [1]. The continuing Internet revolution has likewise resulted in the global popularity of social media. The first social media site (i.e. Six Degrees) was created in 1997 before blogging sites became popular in 1999 [2]. Kaplan and Haenlein defined social media as 'a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content' [3]. People use social media to connect to one another to share expertise, knowledge and opinion amongst social media users; exchange information or ideas; increase sales of products and as medium for teaching, amongst others [4]–[7]. On a positive note, social media has become a popular socialising tool because people created networks without being limited by time and distance [8]. Unfortunately, the advantages of social media are occasionally negated when some people engage in illegal online activities or cybercrime, such as cyberbullying [9].

1.1 Definition of Cyberbullying

Cyberbullying has been given several definitions, including those related to cyberstalking [10]. Cyberstalking is defined as 'stalking other people information to make a false accusation, monitoring, identity theft, threats, and create data destruction or manipulation' [11]. However, some scholars have classified cyberstalking as a type of cyberbullying [8], [9], [11]–[13]. Cyberbullying has emerged with the rapid development of social networks [14]. Accordingly, cyberbullying has been acknowledged as a serious national health threat, and the US Centers for Disease Control and Prevention (CDC) have warned the general public on this issue [10], [15]. Although studies related to cyberbullying have provided varying definitions, the following elements have been identified as common characteristics of cyberbullying: use of electronic devices, intent to cause harm, anonymity and publicity [14, 16]. Anonymity refers to situations when victims have no idea about the identity of harassers, thereby possibly heightening feelings of powerlessness and frustration [17]. Publicity indicates that cyberbullying happens in public social networking sites, in which videos, pictures or messages are distributed publicly. A previous study has indicated that students admit that the most severe cases of cyberbullying pertain to incidents involving a large and public audience [17].

Although the definition of cyberbullying is similar to that of traditional bullying, the former can be referred to as bullying by using the medium of social networks [18]. Slonje and Smith defined cyberbullying as ‘an aggressive, intentional act or behaviour that is carried out by a group or an individual repeatedly and over time through modern technological devices such as mobile phones or internet, against a victim who cannot easily defend him or herself’ [17]. A. Saravanaraj et al. stated that cyberbullying involves an individual or a group of social networking users utilising information and communication technology (ICT) with the intent to harass other users [19]. Patchin and Hinduja defined cyberbullying as ‘willful and repeated harm inflicted through the medium of electronic text’ [20]. Cyberbullying has also been defined as ‘an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself’ [21]. Harassers involved in cyberbullying refer to people who bully victims via online social networks, whilst victims refers to those targeted by harassers in cyberbullying situations.

The majority of researchers have used the preceding definitions as bases in recognising cyberbullying using three main criteria: (i) harassers’ intent towards victims, (ii) repetition of cyberbullying incidents and (iii) strength imbalance between harassers and victims. Firstly, cyberbullying is different from face-to-face bullying. Hence, harassers’ intention towards victims is difficult to interpret because the related incidents do not show the former’s facial expressions (e.g. angry, cynical), voice intonation and body language. Moreover, harassers can take exploit using social media to hide their intentions. By contrast, victims may misinterpret the true intentions of harassers [22]. Secondly, cyberbullying incidents happen repetitively and severely as harassers could bully other victims. Repetitive cyberbullying can have negative implications on victims. For example, if harassers spread fake news related to the victims to other social network users, causing embarrassment on the victims’ part, then this type of cyberbullying can be categorised as defamation [23]. Generally, when rumours are spread in a telecommunication medium, it is difficult to delete and can be accessed by anyone [24]. Thus, before cyberbullying worsens, appropriate actions should be taken, such as automatically detecting cyberbullying in social networks. Thirdly, imbalance between harassers and victims is difficult to determine when bullying occurs in a telecommunication medium. For example, bullying that happens face-to-face may involve harassers being more physically imposing (e.g. taller, more muscular) than victims. Hence, physical imbalance is evident. However, imbalance in cyberbullying, particularly in social media, refers to scenarios involving harassers with superior technological skills or having the ability to manipulate their identities. Consequently, victims may be unable to escape social media harassers [24].

1.2 Consequences of Cyberbullying

Shafie et al. identified two types of bullying effects (i.e. short- and long-term effects) on victim [25]. Short-term effects can be seen from the psychological aspect, such as regular absence in school, deteriorating academic performance and attempts at self-harm. In terms of physical effects, victims may experience *corpus callosum*, which is an abnormal condition in the brain that affects vision and memory [26]. Additionally, victims can experience somatic problems, such as dizziness, headaches, stomach aches and flu. Long-term effects involve psychological, physical, learning, career and social life effects. These effects can be seen when victims become adults with low academic achievements and having difficulty managing finances.

Given the preceding negative consequences, cyberbullying detection has become one of the predominant solutions to cybercrime. Thus, studies on natural language processing (NLP), such as Al-Garadi et al. (2016), Zhao et al. (2016), Van Hee et al. (2018) and Talpur and O’Sullivan (2020), have been conducted to solve this problem in social networks [27]–[30]. Despite the extensive research, cyberbullying remains prevalent amongst social network users. The reason is that cyberbullying in social networks involving computational linguistics is difficult to identify because this crime occurs in a variety of ways in different social networks [28]. Evidently, this matter has become one of the challenges in the field of automated cyberbullying detection.

1.3 Background of the Study

The current study proposed an improvement in developing a cyberbullying detection model, which focuses on the social media platform ASKfm, by using machine learning. This research adopted machine learning to train the cyberbullying detection model based on existing data set to predict whether or not a new data set consists of indicators of cyberbullying. Supervised learning (instead of unsupervised or reinforcement learning) was implemented because the data set was completely labelled in a previous study [28]. To train the model, supervised learning, which is a machine learning task, learns a function that maps input to output based on a labelled data set to predict outcomes accurately [31]. We extracted content-based features to be fed into an algorithm to develop a predictive model. We implemented a linear support vector classifier (SVC). Given our focus on supervised learning, corpus from the social media platform ASKfm was used as our data set annotated into two classes: cyberbullying (labelled as 1) and non-cyberbullying (labelled as 0) [28]. ASKfm, which is a popular Latvian social media platform created in 2010 in Riga, Latvia, is a website that enables users to ask any questions without the need to reveal their respective identities (anonymous) [32]. We implemented an over-sampling technique called Synthetic Minority Over-sampling TEchnique (SMOTE) proposed by Chawla et al. [33]. Hyperparameters of linear SVC was also implemented to develop the most efficient model. The current study proposed the following cyberbullying detection model framework:

- We proposed four significant content-based features: term frequency–inverse document frequency (TF-IDF) for word and character, pre-trained word embedding using Word2vec and six types of term list: profane words, proper nouns, negation words, ‘allness’ term, diminisher words and intensifier words.

- We trained different feature combinations iteratively using linear SVC in four environments: default environment, tuned hyperparameter environment only, using SMOTE environment only and a combination of hyperparameter tuning with SMOTE environment.

The remainder of this paper is organised as follows. Section 2 presents the related research on cyberbullying detection. Section 3 elaborates the methodology. Section 4 reports the result of the exploratory data and experimental analyses. Lastly, Section 5 discusses the experimental result and provides the conclusion.

2.0 RELATED RESEARCH

Cyberbullying could happen anywhere and anytime, and this issue is closely related to the safety of people [34]. New technology, such as Web 3.0, has popularized social networks, particularly amongst teenagers, and empowered users to become anonymous individuals whilst in cyberspace. Additionally, behaviours deviating from normal user habits in social networks are considered part of cybercrimes and can be categorised as cyberbullying [35]. To overcome the challenge of detecting cyberbullying in social networks, cyberbullying detection models using machine learning, deep learning, neural network and NLP have been given serious attention by researchers involved in the security of machines [34]. Sources from social media, such as comments, posts, videos or pictures, could be input for data analytics and NLP fields and could provide insights into topics of interest [36]. The use of features and classifiers has a significant impact on assessing a cyberbullying detection model.

This study used machine learning research because of its success in producing numerous models, tools and algorithm when handling a large data set when solving real-world problems [37], [38]. The objective of machine learning is to define and identify correlations and patterns of data. Accordingly, model performance is improved by describing meaningful correlations and patterns in a training data set to gain knowledge from experience before feeding the model with new patterns and correlations of a data set when predicting unseen data [34]. When this concept is implemented in supervised machine learning, such as for classification use case, classification task is learned with the help of the relevant training data set.

2.1 Features in Cyberbullying Detection

Content-based features are the most widely used features in cyberbullying detection studies in the context of linguistic computerisation. Social network users utilise text-based messages to convey cyberbullying; unsurprisingly, profane words have been found to be the most widely used feature in cyberbullying detection [39]. Dinakar et al., Kontostathis et al. and Nahar et al. produced lexical profane words using external sources, such as noswearing.com[1] and

[1] https://www.noswearing.com/

urbandictionary.com[2] [40]–[42]. However, implementing only profane words as a single feature in cyberbullying detection is incapable of considering messages contained in cyberbullying incidents. Yin et al. noted that the use of profane words alone does not provide high retrieval (Recall) [43]. The use of cyberbullying keywords as a feature is also widely used in cyberbullying detection studies of Mahmud et al., Dinakar et al. and Dadvar et al. [40], [44], [45]. Word embedding became popular in cyberbullying detection by implementing Word2vec [46] [47] [48]. The two architectures of Word2vec word embedding are continuous bag of words (CBOW) and skip-gram [49]. Several studies related to cyberbullying detection have also implemented TF–IDF [40], [43], [50].

2.2 Cyberbullying Detection Using Machine Learning

The majority of cyberbullying detection methods have used machine learning techniques to detect cyberbullying. Yin et al. (2009) was the earliest to use this technique to incorporate TF–IDF, n-grams, pronoun, profane words and polarity as features to train linear kernel SVM [43]. They used three data sets to compare the model performances of Kongregate, Slashdot and Myspace. The result from the experiments described TF–IDF is one of the features with the most contributions in the model performance when combined with other features.

Chavan and Shylaja [51] implemented machine learning to identify insults and offensiveness in comments on cyberbullying incidents in social networking sites. They extracted the following features from the Kaggle data set: n-grams, frequency of words, TF–IDF, occurrence of pronouns and skip-grams. Additionally, they implemented feature selection chi-square to handle numerous features. They likewise adopted two algorithms, namely, SVM and logistic regression, in their experiments. The logistic regression achieved 73.76% accuracy, 60% recall and 64.4% precision. SVM achieved 77.65% accuracy, 58% recall and 70% precision. Perera and Fernando [52] adopted supervised machine learning to detect cyberbullying, specifically by using labelled data set from Twitter. They extracted TF–IDF, sentiment analysis, profanity and pronoun to be fed into SVM as classifier in their research. Their experimental result indicated that the combination of TF–IDF and sentiment features provides the best result amongst others in terms of accuracy, precision, recall and FI-measure.

R. Shah et al. [53] implemented machine learning to detect cyberbullying content in Twitter, and adopted five classifiers: SVC, logistic regression, multinomial naïve Bayes, random forest and stochastic gradient descent (GSD). Word TF–IDF was the only feature used in their study. Their result indicated that logistic regression provides accurate classification with 91% precision, 94% recall and 93% FI-measure.

Van Hee et al. [28] extracted five features for implementation in Linear SVC LIBLINEAR: word n-grams (i.e. unigrams, bigrams and trigrams), character n-grams, subjectivity lexicon, term lists and topic models. Word n-grams (without hyperparameters tuning) and profane words were the baseline of features in this study. Data set from ASKfm was collected and annotated using the brat rapid annotation tool (BRAT). A total of 31 combinations of features were on 28 hyperparameters sets, and models were evaluated using F-measure. This study will be our benchmark.

TF–IDF, sentiment and n-grams were extracted as features to be trained using SVM and neural network without implementation of the resampling technique [54]. Additionally, n-grams were divided into bigrams, trigrams and fourgrams. Polarity of sentiment features were extracted using Text Blob library and combined with TF–IDF before training the model using the Kaggle data set.

2.3 Cyberbullying Detection Using Deep Learning and Neural Network

Apart from machine learning, other techniques that have been used recently in cyberbullying detection are deep learning and neural network. Zhang et al. [55] proposed a novel pronunciation-based convolutional neural network (PCNN) to detect cyberbullying in 1313 messages from Twitter and 13 000 messages from Formspring.me. PCNN achieved the best result for Formspring.me compared with the two CNN baselines, with 74.00% precision, 45.30% recall, 96.80% accuracy and 56.20 F1-measure. PCNN also outperformed all models, such as random forest, SVM, multilayer perceptron, J48 decision tree, CNN pre-trained and CNN random. However, precision was slightly less compared with CNN random, with differences of approximately 0.3%, in which the Twitter data set resulted in 99.10% accuracy, 97.0% recall, 98.90% accuracy and 98.00% F1-measure.

Al-Ajlan and Ykhlef [56] adopted deep learning to detect cyberbullying and proposed a novel algorithm called CNN-CB that removes feature engineering. CNN-CB is based on CNN. Data set from Twitter was used to perform

[2] https://www.urbandictionary.com/

experiments. The results indicated that the proposed method outperformed SVM algorithm, with 95% accuracy obtained by CNN-CB.

Iwendi et al. [57] performed an empirical analysis to determine the performance and effectiveness of deep learning algorithms in cyberbullying detection. Four deep learning model were adopted in their experiments: bidirectional long short-term memory (BLSTM), gated recurrent units (GRU), LSTM, and recurrent neural network (RNN). They implemented a data set from the Kaggle repository. The experimental results indicated that BLSTM model performed better than RNN, LSTM and GRU in terms of accuracy (82.18%) and F1-measure (88%).

The use of deep learning and neural network in cyberbullying detection is beneficial in effectively detecting cyberbullying in social media. Nonetheless, we decided to adopt machine learning technique because deep learning is computationally expensive to train and complex models may take time to train.

3.0 METHODOLOGY

A few processes should be done in this study to develop a cyberbullying detection model in ASKfm. This section explains the process and experiment setting. Fig. 1 shows the proposed cyberbullying detection framework.

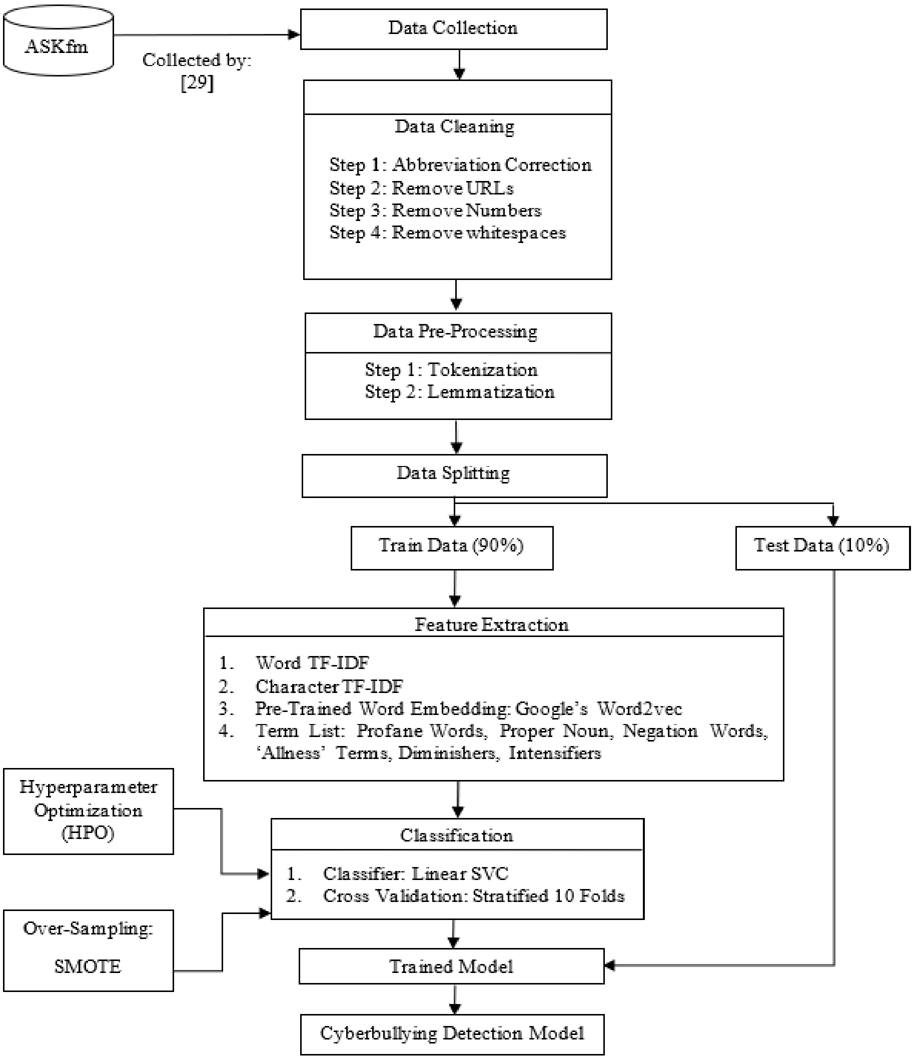

ASKfm
Collected by: [29]
Data Collection
Data Cleaning
Step 1: Abbreviation Correction
Step 2: Remove URLs
Step 3: Remove Numbers
Step 4: Remove whitespaces
Data Pre-Processing
Step 1: Tokenization
Step 2: Lemmatization
Data Splitting
Train Data (90%)
Test Data (10%)
Feature Extraction
1. Word TF-IDF
2. Character TF-IDF
3. Pre-Trained Word Embedding: Google's Word2vec
4. Term List: Profane Words, Proper Noun, Negation Words, 'Allness' Terms, Diminishers, Intensifiers
Hyperparameter Optimization (HPO)
Over-Sampling: SMOTE
Classification
1. Classifier: Linear SVC
2. Cross Validation: Stratified 10 Folds
Trained Model
Cyberbullying Detection Model

Fig. 1: Proposed framework of cyberbullying detection model using the data set from ASKfm

3.1 Data Collection

This study used a secondary data set called AMiCA Bullying Cyber Dataset, which was used by [28]. This data set was collected from ASKfm from April 2013 to October 2013. Thereafter, the authors [28] turned over the data set to annotators to proceed with annotation for each message in the data set using BRAT [58]. Therefore, the data set we received are in two formats: (1) text file consisting of original conversation from ASKfm users and (2) annotation file comprising messages that have been annotated. Our task is to parse both files and convert them into an eXtensible Markup Language (xml) file format to accommodate the implementation during the experiment. Accordingly, we used BratReader and altered the coding based on the requirements of our study. BratReader is a Python code for reading BRAT repositories and converting text or annotation file format into xml for easy access [59]. Fig. 2 shows the flow chart of preparing the data set from the text and annotation files into an xml format before we convert into comma-separated values (csv) file for data cleaning. Lastly, the data set consisted of 113 021 messages (posts) after undergoing the data preparation process, as shown in Fig 3. Messages that declared or have indication of cyberbullying were annotated as cyberbullying and those without indication of cyberbullying were annotated as non-cyberbullying.

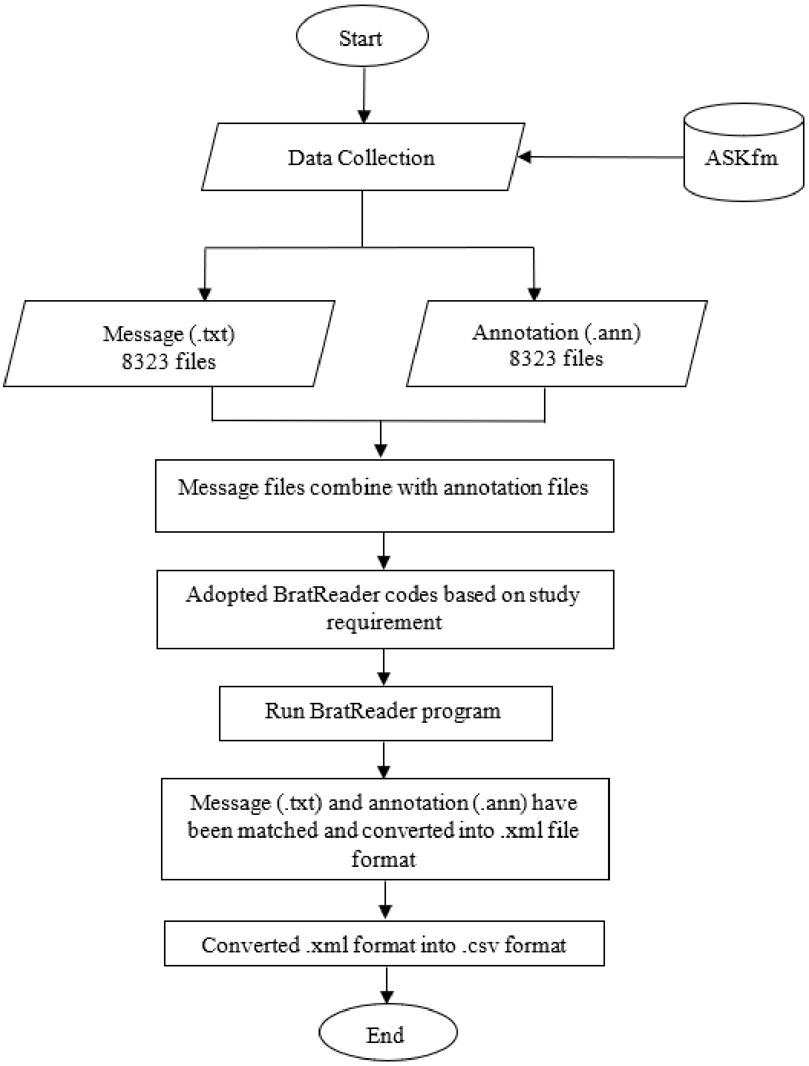
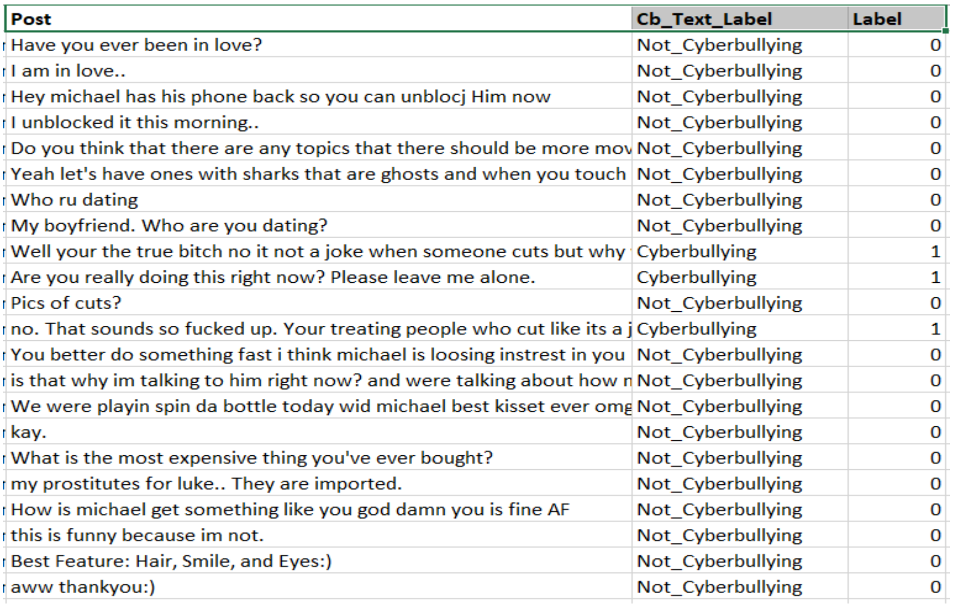

Start
Data Collection
ASKfm
Message (.txt)
8323 files
Annotation (.ann)
8323 files
Message files combine with annotation files
Adopted BratReader codes based on study requirement
Run BratReader program
Message (.txt) and annotation (.ann) have been matched and converted into .xml file format
Converted .xml format into .csv format
End

Fig. 2: Preparation of the data collection into .xml format



Post	Cb_Text_Label	Label
Have you ever been in love?	Not_Cyberbullying	0
I am in love..	Not_Cyberbullying	0
Hey michael has his phone back so you can unblocj Him now	Not_Cyberbullying	0
I unblocked it this morning..	Not_Cyberbullying	0
Do you think that there are any topics that there should be more mov	Not_Cyberbullying	0
Yeah let's have ones with sharks that are ghosts and when you touch	Not_Cyberbullying	0
Who ru dating	Not_Cyberbullying	0
My boyfriend. Who are you dating?	Not_Cyberbullying	0
Well your the true bitch no it not a joke when someone cuts but why	Cyberbullying	1
Are you really doing this right now? Please leave me alone.	Cyberbullying	1
Pics of cuts?	Not_Cyberbullying	0
no. That sounds so fucked up. Your treating people who cut like its a j	Cyberbullying	1
You better do something fast i think michael is loosing instrest in you	Not_Cyberbullying	0
is that why im talking to him right now? and were talking about how n	Not_Cyberbullying	0
We were playin spin da bottle today wid michael best kisset ever omg	Not_Cyberbullying	0
kay.	Not_Cyberbullying	0
What is the most expensive thing you've ever bought?	Not_Cyberbullying	0
my prostitutes for luke.. They are imported.	Not_Cyberbullying	0
How is michael get something like you god damn you is fine AF	Not_Cyberbullying	0
this is funny because im not.	Not_Cyberbullying	0
Best Feature: Hair, Smile, and Eyes:)	Not_Cyberbullying	0
aww thankyou:)	Not_Cyberbullying	0

Fig. 3: Screenshot of the ASKfm data set

3.2 Data Cleaning

Data cleaning is one of the important components of modelling the predictive model by using machine learning. This process refines our data quality by removing noise, errors and inconsistent data in the data set [60], [61]. We performed data cleaning prior to conducting data pre-processing process. The four steps involved in this process are shown in Fig. 1, in which the first step was adopted from [62] to detect misspelled words and implement corrections by mapping the abbreviation dictionary. The abbreviation dictionary was compiled from two external sources; Netlingo[3] and Slangit[4]. Thereafter, uniform resource locator (URLs), numbers and whitespaces were removed from the data set.

3.3 Data Pre-Processing

Prior to the subsequent data splitting, data pre-processing should be conducted by performing two important steps in text processing: tokenisation and lemmatisation. Library for machine learning in Python (i.e. scikit-learn) was used to integrate algorithms, and natural language toolkit (NLTK) was utilised to process the data set. We used NLTK to implement word_tokenize for splitting messages into each word. After all words have been tokenised, lemmatisation followed to find the lemma of each word according to its context and meaning by implementing NLTK WordNetLemmatizer. To enhance our understanding, we visualised the steps for data pre-processing in Fig. 4. Table1 shows an example of data pre-processing involving tokenisation and lemmatisation.

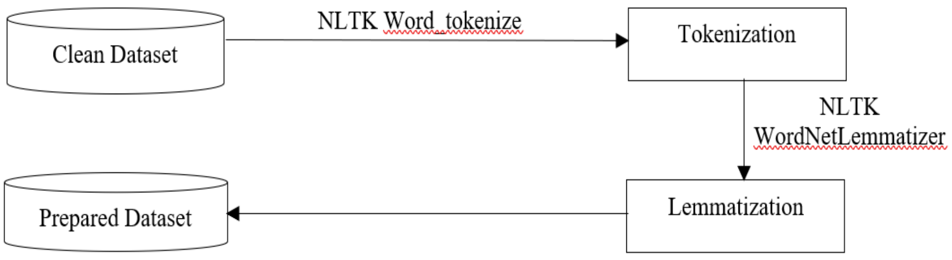


Fig. 4: Visualisation of data pre-processing steps that occurs in the scikit-learn library

[3] https://www.netlingo.com/acronyms.php
[4] https://slangit.com/terms/social_media-all

Table 1: Example of output for data pre-processing

Index	Post	Tokenization	Lemmatization
1	Im going to bed	[‘im’, ‘going’, ‘to’, ‘bed’]	[‘im’, ‘go’, ‘to’, ‘bed’]
2	I actually cant even cried	[‘I’, ‘actually’, ‘cant’, ‘even’, ‘cried’]	[‘I’, ‘actually’, ‘cant’, ‘even’, ‘cry’]
3	My boyfriend Who are you dating	[‘My’, ‘boyfriend’, ‘Who’, ‘are’, ‘you’, ‘dating’]	[‘My’, ‘boyfriend’, ‘Who’, ‘be’, ‘you’, ‘date’]
4	What is the most expensive thing you’ve ever bought?	[‘What’, ‘is’, ‘the’, ‘most’, ‘expensive’, ‘thing’, ‘you’, ‘’ve’, ‘bought’, ‘?’]	[‘What’, ‘is’, ‘the’, ‘most’, ‘expensive’, ‘thing’, ‘you’, ‘’ve’, ‘buy’, ‘?’]
5	Name any things that make you smile?	[‘Name’, ‘any’, ‘things’, ‘that’, ‘make’, ‘you’, ‘smile’, ‘?’]	[‘Name’, ‘any’, ‘things’, ‘that’, ‘make’, ‘you’, ‘smile’, ‘?’]

3.4 Data Splitting

In machine learning, data are split into two parts using rule of thumb: train and test data sets. This study split the data into 90% train data and 10% test data set by implementing train–test split in scikit-learn for evaluating the linear SVC performance. Given that the data set used in this experiment is secondary data set, which were collected and implemented by [28], we follow the data splitting ratios from the aforementioned research.

3.5 Feature Extraction

Feature extraction is one of the crucial parts in machine learning modelling because the performance of classifier depends on features used during the classification process [63]. The current study chose four types of features in the category of content-based features only for a comparison of the model performance evaluation before we proceed and integrate another feature category (e.g. sentiment or user feature in a future study). Feature will be extracted from raw data set, converted into a vector of token and represented in matrix form. The total features extracted is 12,095. The following features were extracted after pre-processing in this study.

1. **Word TF–IDF:** TF–IDF is important in measuring a word to a document in a data set. Word TF–IDF is amongst the features that have been used in cyberbullying detection since 2009. On the basis of state-of-the-art cyberbullying detection, word TF–IDF could enhance the performance of the cyberbullying detection model. Thus, we extracted 3131 TF-IDF in word-level feature in n-grams (i.e. ranging from unigrams, bigrams to trigrams) to calculate the weighting terms by importance.

2. **Character TF–IDF:** As word TF–IDF, 5714 TF–IDF character-level features were extracted with crossing word boundaries (char_wb) in n-grams ranging from bigrams, trigrams to fourgrams. Tfidfvectorizer in scikit-learn was used to encode the data set into the TF–IDF matrix. Character level in TF–IDF weighting is one of the features in the current study because it could contribute a good percentage in evaluation performance based on spelling diversification [28]. This feature was also used to determine the arrangement between characters because current social media users constantly use word abbreviations in their writing style.

3. **Word embedding:** In NLP tasks, machine learning algorithms depend on word embedding. Thus, we also made word embedding one of the features to convert the text for each message in the data set to numeric form [63]. We used pre-trained Word2vec[5] for word embedding feature extraction with 300 dimensions. Features were extracted by calculating the vector average of words in each post. We calculated that each

[5] https://code.google.com/archive/p/word2vec/

message Word2vec is one of the word embedding techniques for NLP published in 2013, and used a neural network model to learn word from a corpus [64].

4. **Term list:** By adopting features used in [28], [65], we extracted six categories of features called term lists, which are derived from one binary feature. All redundant terms were discarded and only 2950 unique terms were listed.

    - **Profane words:** We used two external sources consisting of curse word lexicon[6] and Google profanity list[7] for the assembled profane words in a dictionary. Curse word lexicon and Google profanity list contain 1383 and 458 words, respectively. After removing redundant words, our final dictionary contained 1681 words.
    - **Proper noun:** By using NLTK called entity recognition, we extracted the features of proper nouns contained in the ASKfm data set. The total number of proper nouns is approximately 974.
    - **Negation words:** The use of negation words, such as *no, without, never* or *nothing*, in a sentence may alter the direction and meaning of such a sentence. Hence, we included negation words as our feature in cyberbullying detection. We assembled negation words according to [66], [67], and the total number of negation words listed was approximately 42 words.
    - **'Allness' terms:** According to [68], people constantly express their emotions through speaking and writing to attract attention from others. They tend to use terms with no limits, such as *always, everything, completely* and *no more*. Accordingly, we extracted 'allness' terms to detect cyberbullying in ASKfm, and 210 terms were collected from two external sources and one article: List of Absolutes[8], Grammar: Absolute Words[9] and [69].
    - **Diminishers:** Diminisher is a term that could be used to turn the direction of sentiment in a sentence. Example of diminisher terms are *bit*, *few*, *quiet* and *some*. We implemented 35 terms of diminisher from [67] as our features to be adopted in the experiments.
    - **Intensifiers:** In linguistics, an intensifier is one of the lexical categories and an adverb, such as *terribly*, *ridiculously*, *pretty* and *rather*. It highlights, intensifies and down tones another adverb, verb or adjective in any sentence [70]. Thus, we compiled 82 terms of intensifier from One Minute English[10], Learn English British Council[11] and [67].

3.6 Classification

SVM is one of the supervised learning models in machine learning and amongst the most popular algorithms used for classification in detection as task for NLP. SVM learns the algorithms to analyse data, works well with high-scale dimensional data and is sensitive to hyperparameter optimisation because it is able to change abruptly whilst working [39][71]. SVM is one of the algorithms in machine learning technique, such as classification, and is suitable for cyberbullying detection to predict whether or not the text used in ASKfm contained cyberbullying. When we extracted features from the text, it will turn into high-scale dimensional data depending on the data set size. SVM also performs exceptionally under different conditions, such as types of feature used and percentage of missing data [39]. State-of-the-art cyberbullying detection shows that SVM is one of the best classifiers used in experiments and is able to develop efficient predictive model [39]. Furthermore, experiments in the current study aim to classify two classes between cyberbullying and non-cyberbullying. To perform binary classification in our experiments, we used linear SVC, which we implemented based on LIBLINEAR [28], [72]. Linear SVC was chosen as classifier in this study because it performs faster compared with linear kernel and Gaussian kernel, which were implemented in LIBSVM. Thus, the advantage of linear SVC is its speed of training and predicting data [73]. Linear SVC is one of the functions in the SVM module in scikit-learn, and fits the data by finding the best linear hyperplane separating classes with the maximum margin, as shown in Fig. 5 [73].

[6] https://www.cs.cmu.edu/~biglou/resources/bad-words.txt

[7] https://code.google.com/archive/p/badwordslist/downloads

[8] http://nomistakespublishing.com/writing-resources/list-of-absolutes/

[9] https://kddidit.com/2015/04/20/grammar-absolute-words/

[10] https://oneminuteenglish.org/en/list-intensifiers/

[11] https://learnenglish.britishcouncil.org/english-grammar-reference/intensifiers

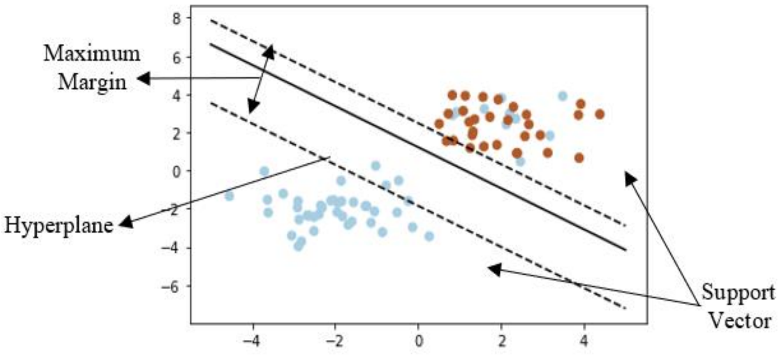


Fig. 5: 2D plot of linear SVC using Python

3.7 Imbalance Class Distribution

Fig. 6 shows the distribution of the ASKfm data set using a statistical graph. As shown in Fig. 6, our final data set is approximately 113 021 (i.e. data set annotated as cyberbullying, 5375; and non-cyberbullying, 107 646). Evidently, the data set is imbalanced with the percentages of cyberbullying and non-cyberbullying approximately 4.76% and 95.24%, respectively, which is a moderate imbalanced data set. The annotation percentage of the ASKfm data set is shown in Fig. 7.

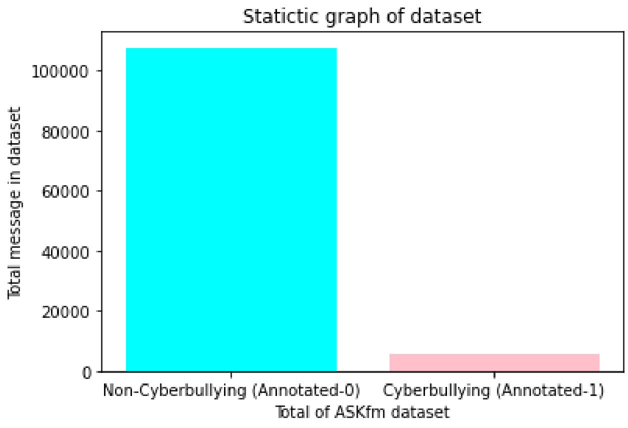


Fig. 6: Distribution of the ASKfm data set

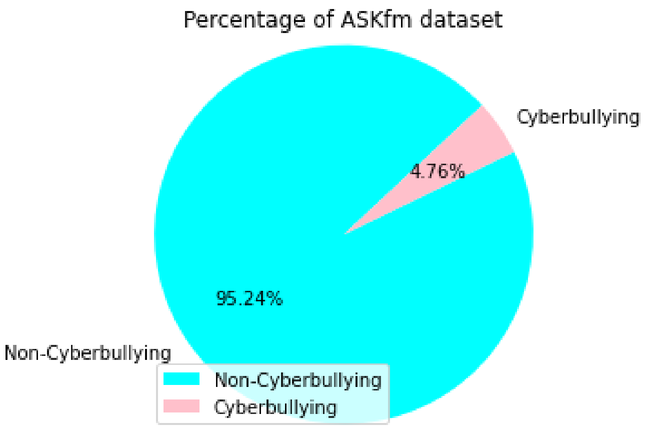


Fig. 7: Percentage of distribution of the ASKfm data set

Real-world data are constantly imbalanced for classification problem, such as spam, fraud detection and churn detections. Data provided to algorithms are typically imbalanced and lead to poor model performance evaluation. Imbalanced distribution of data means class samples are biased or skewed [33], [74]. Given that our data set is biased between two classes, namely, cyberbullying (annotated as 1) and non-cyberbullying (annotated as 0), we implemented over-sampling SMOTE. This selection is based on the cyberbullying detection experiments conducted by Al-Garadi et al. [27], showing that over-sampling SMOTE improves classifier performance. However, under-sampling was not used in our study because this technique down sized the major class randomly, which may risk eliminating the important data set [75]. SMOTE generates synthetic minority class samples in the training data set before training the algorithm. Table 2 shows the training data before and after applying SMOTE.

Table 2: Training data before and after applying SMOTE

SMOTE	Data Annotation	Total Training Data
Before Over-Sampling	1 (cyberbullying) 0 (non-cyberbullying	4820 96 898
After Over-Sampling	1 (cyberbullying) 0 (non-cyberbullying	96 898 96 898

3.8 Hyperparameter Optimisation

Fundamental task in machine learning system is setting hyperparameter to optimise the performance of the developed model [76], [77]. Technically, hyperparameter tunes machine learning model using the training data to obtain a set of hyperparameter to determine the minimum error on testing data. Hyperparameter optimisation is important because the implementation of hyperparameter tuning results in an effortless development of machine learning models [76]. Prediction model performance can be improved with the hyperparameter tuning model [77]. Grid search is a well-known method for fine-tuning the best parameters for various machine learning algorithms. Particularly, grid search uses cross-validation to guide the performance metrics. Grid search is an exhaustive search that can be used to find the best hyperparameter values. It can create a model that generates each parameter combination whilst also storing each model combination. This grid search can help save time and resources. Thereafter, the classifier models are built using the tuned parameters. Test data are applied to the proposed model with the tune hyperparameter once the classification model is generated before the test data performance is evaluated. Instead of blind search, it requires a lengthy process and large memory. Each iteration in informed search learns from the previous one. However, all modellings in grid search are done simultaneously, in which the best one is chosen. Grid search would be faster compared with informed and blind searches. The current study performed the hyperparameter of linear SVC in grid search for the comparison in the experiments. Grid search used library function from scikit-learn GridSearchCV to help in the loop through the predefined hyperparameters and fit the estimator on the training data set, resulting in the best hyperparameters of linear SVC. [78]. Table 3 lists the hyperparameters of the linear SVC used.

Table 3: Hyperparameters of linear SVC

Key	Description	Value
C	Error term penalty	0.1,1
Class_weight	Set penalty C of class i	balanced
Max_iter	Iteration number that needs to be ran	120 000
Loss	Loss function where ‘hinge’ is standard SVM loss and ‘squared_hinge’ is square of hinge loss	hinge, squared_hinge
Multi_class	For better accuracy	Ovr, crammer_singer

3.9 Experimental Setup

We ran large-scale sets of experiments to evaluate the linear SVC performance towards the development of a cyberbullying detection model via ASKfm. Linear SVC was trained and tested in different experimental settings. Firstly, the experiments were conducted in four settings: default, hyperparameter optimisation, SMOTE and hyperparameter + SMOTE. Secondly, we mapped each feature to indicator as shown in Table 4. Thereafter, the power sets (except for the empty set) of the four features {A, B, C, D} were used to train the linear SVC. Therefore, 60 training and test trials were performed. In Subsection 4.2, the results of all feature sets will not be reported owing to space limitation. Amongst the feature sets consisting of a single feature (i.e. {A}, {B}, {C}, {D}), only the result of the best one will be reported. Similarly, we will report the results of the best feature sets amongst the combination of two, three and four (all) types of features. All experiments were conducted in a controlled environment, in which all data were analysed on a single machine to ensure consistency in the results. The configurations of the hardware and software are presented in Table 5.

Table 4: Feature mapping for Tables 7, 8, 9 and 10

Feature	Indicator
Word TF-IDF	A
Character TF-IDF	B
Word Embedding	C
Term List	D

Table 5: Information of hardware and software components used in the current study

Component	Information
Random Access Memory (RAM)	Intel Core i7
Central Processing Unit (CPU)	24GB
Operating System (OS)	64-bit Window 10
Python	Version 3.8

4.0 RESULTS

This section presents two results of our exploratory data and experimental analyses of cyberbullying detection. Exploratory data analysis elaborated data set distribution, whilst experimental analysis presented the experiments conducted as explained in the experimental setup.

4.1 Exploratory Data Analysis

W. Tukey explained that the definition of data analysis includes several important aspects involving procedures in analysing data, techniques for interpreting the results of the procedure, data collection planning to obtain accurate results and involvement of statistical mathematics [79]. Therefore, the first step before starting any experiment is to understand the characteristics of the data used. This study is known as exploratory data analysis, which is a data science process involving the conduct of a preliminary investigation of the collected data. Exploratory data analysis is also one of the methods in manipulating data sources to obtain answers for the research conducted.

A word cloud is a visual presentation of text processing to show a collection or cluster of specific word frequencies used in a corpus, in which words with high usage frequencies will be labelled as large and presented in boldface with different colours [80]. Tessem et al. stated that the smaller the displayed lexicons, the less frequency that a word is used [80]. Fig. 8 presents a visualisation of a word cloud for a data set annotated as cyberbullying, in which the majority of the words were profane words, such as *fuck*, *assno*, *fucking*, *bitch*, and *hate*, used during conversations in ASKfm. Word frequency is visualised in the form of a bar graph, as shown in Fig. 9. Additionally, Fig. 9 shows that the majority of the word usage frequencies found in the 5375 messages annotated as cyberbullying were profane words, such as *fuck* (864), *assno* (479), *fucking* (393), *bitch* (366), *ugly* (287), *hate* (282), *shit* (275), *cunt* (196), *ass* (166), *slut* (165), *fat* (164), *dick* (160), *stupid* (128) and *gay* (104). Additionally, there were also words that are not profane words, including *people* (309), *leave* (187), *shut* (162), *life* (155), *fake* (134), *question* (125) and *love* (115). This analysis further indicated that profane words can be used as one of the features in this study to classify classes and build models for cyberbullying detection.

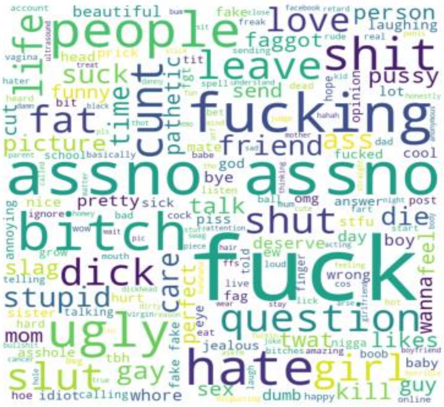


Fig. 8: Word cloud visualisation of the ASKfm data set annotated as 1 (cyberbullying)

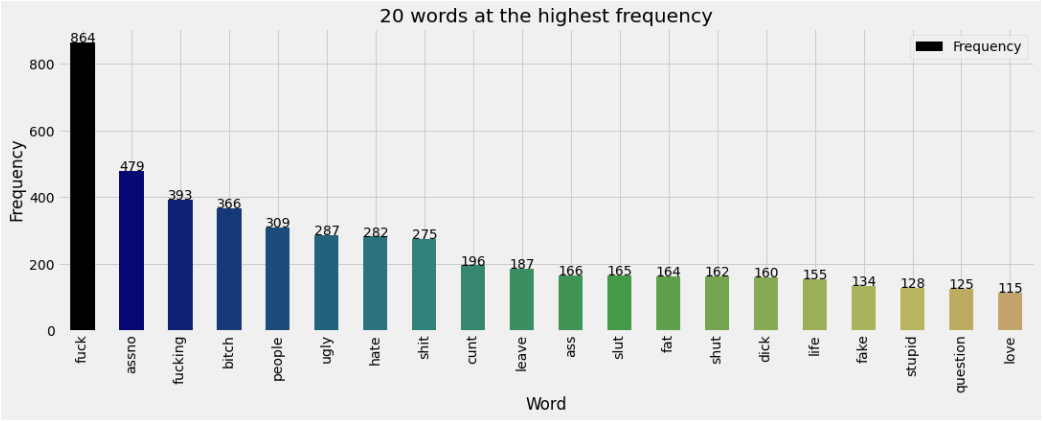


Fig. 9: Twenty of the most frequent words in the ASKfm data set annotated as 1 (cyberbullying)

Fig. 10 shows a word cloud based on 107 646 messages that do not indicate cyberbullying. *Love*, *question*, *likes*, *people* and *reply* are amongst the words often used. Profane words, such as *fuck* and *fucking*, are also amongst the words displayed in the word cloud. However, these lexicons were not frequently used by ASKfm users in the data set. As shown in Fig. 10, *love* (3757) is the highest frequency word, as visualised in Fig. 11, followed by *question* (3499), *likes* (3014), *people* (2323), *reply* (1917), *follow* (1560), *person* (1533), *beautiful* (1481), *pretty* (1479), *nice* (1409), *time* (1291), *girl* (1284), *life* (1274), *idk* (1192), *friends* (1169), *perfect* (1164), *day* (1127), *hot* (1126), *favourite* (1102), *cute* (1073) and *talk* (1047).

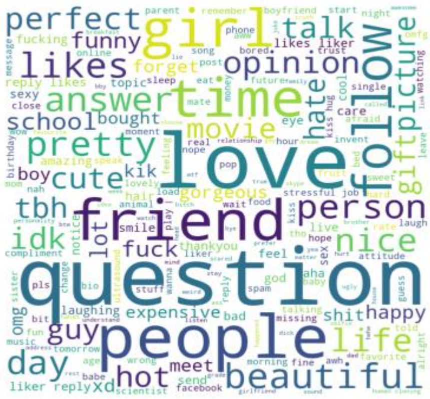


Fig. 10: Word cloud visualisation in ASKfm data set annotated as 0 (non-cyberbullying)

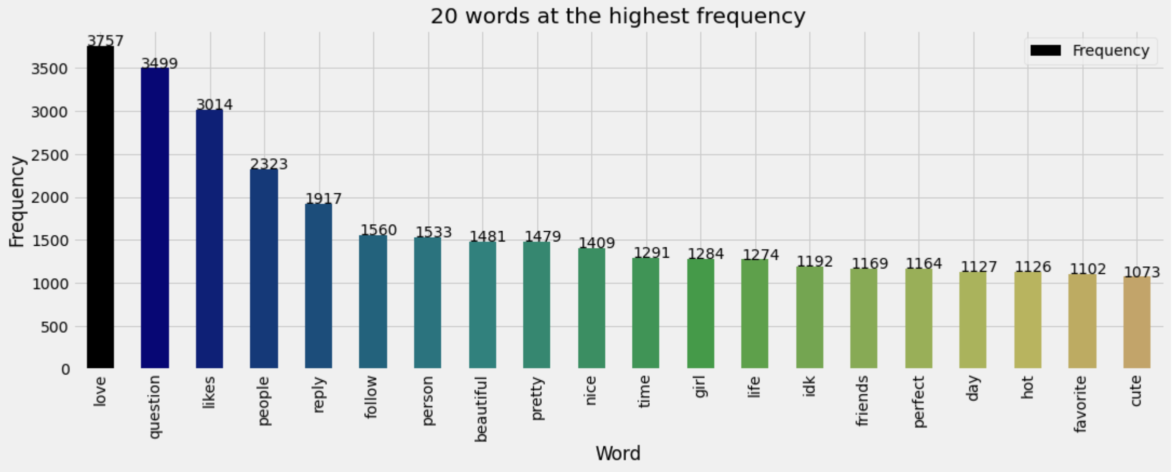


Fig. 11: Twenty most frequent words in ASKfm data set annotated as 0 (non-cyberbullying)

4.2 Experimental Analysis

Experiments have been running over a 10-fold cross-validation in grid search with different settings. Our data set contained imbalanced class distribution between cyberbullying and non-cyberbullying. Thus, we considered area under curve (AUC) as our main performance to be evaluated because it is the evaluation metric with impressive robustness for evaluating the developed model. AUC can measure the ability of algorithms to differentiate between classes and as summary for the receiver operating characteristic (ROC) curve. AUC is generally used in medical decision-making, such as evaluating diagnostic tests and predictive model for heart diseases [81], [82]. Additionally, we reported other evaluation metrics, such as accuracy, precision, recall and f-measure, to evaluate the effectiveness of cyberbullying detection models. These metrics gives different calculations on model performance [83]. Table 5 shows a confusion matrix for a cyberbullying detection predictive model that can derive accuracy, precision, recall, f-measure and AUC [84].

Table 6: Confusion matrix of cyberbullying classification

		Predicted	
		Cyberbullying (+)	Non-Cyberbullying (-)
Actual	Cyberbullying (+)	a	c
	Non-Cyberbullying (-)	b	d

Table 5 shows an n × n confusion matrix, where n = 2, with the following meanings for each entry:

- a: Number of positive cyberbullying predictions that are true;
- b: Number of positive cyberbullying predictions that are false;
- c: Number of negative non-cyberbullying predictions that are false and
- d: Number of non-cyberbullying predictions that are true.

Classification accuracy is a metric that measures a classification model’s performance by dividing the number of correct predictions by the total number of predictions. Measurement of accuracy is important because inaccuracies in results may be caused by faulty equipment or poor data processing. In this study, model accuracy can be defined as the ratio of the correct predictions of cyberbullying to the total number of predictions [84]. Mathematically, it can be defined as follows:

Precision is the ratio of cyberbullying predictions that are true to the total number of true cyberbullying and false cyberbullying predictions [85]. That is, precision is a measurement of messages that correctly identifies to contain cyberbullying incidents out of all messages that actually contains cyberbullying incidents. Mathematically, it can be defined as follows:

In binary classification, recall is also known as sensitivity. Recall is the ratio of cyberbullying predictions that are true to the total number of cyberbullying predictions that are true and non-cyberbullying predictions that are false [85]. Recall provides true positive rate (TPR), which is a measure of models correctly identifying true positive. For example, for messages containing cyberbullying incidents, recall provides the number of messages that are correctly identified as containing cyberbullying incidents. Mathematically, it can be defined as follows:

An F-measure is the average of precision and recall [86]. F-measure score ranges from 1 (greatest) to 0 (worst). A low F-measure score indicates a lack of precision and recall. Mathematically, it can be defined as follow:

The AUC value varies from 0 to 1; the greater the AUC, the better. The AUC graph provides the false positive rate (x-axis) versus true positive rate (y-axis) for a variety of candidate threshold values ranging from 0.0 to 1.0. For each conceivable cut-off, the AUC curve depicts the relationship between sensitivity and specificity. The AUC curve is a graph that includes the following information [85]:

4.2.1 Result of Linear SVC with Default Experiments Setting

Linear SVC was used to conduct 15 experiments using all proposed features in default setting (i.e. without tuned hyperparameters and over-sampling SMOTE). Table 7 shows only the highest AUC percentage score for each setting to minimise spaces. For the training data set, the AUC results were between 91.06 and 98.02, with the percentages of the F-measure between 41.05 and 70.33. The best feature group with the highest AUC is 98.02 with an F-measure of 66.22. For the testing data set, the highest AUC percentages between 89.33 to 96.57 and F-measure between 36.12 and 56.21. Thus, the best feature group for default setting is word TF–IDF, character TF–IDF and word embedding. We can hypothesise that these three features have strong features compared with others in the default setting. The testing data set results showed that the trained model generalise well to the new data despite differences between the F-measure and the AUC testing and training data sets at approximately 10.01% and 1.45%, respectively.

Table 7: Performance evaluation of linear SVC for the training and testing data sets (%) according to different metrics (A: accuracy; P: precision; R: recall; F_1: F-measure and AUC: area under the ROC curve) under default experiment setting

Feature	**Training Dataset Score (%)**					**Testing Dataset Score (%)**				
	A	P	R	F_1	AUC	A	P	R	F_1	AUC
One feature: A	96.24	79.94	27.61	41.05	91.06	94.89	75.12	19.62	36.12	89.33
Two features: A + B	97.40	85.93	53.73	66.12	97.97	96.51	74.77	43.78	55.23	93.37
Three features: A + B + C	97.39	85.80	53.92	**66.22**	**98.02**	96.57	75.23	44.86	**56.21**	**96.57**
Four features: A + B + C + D	97.64	87.17	58.94	70.33	97.64	96.47	72.81	44.86	55.52	93.11

4.2.2 Result of Linear SVC with SMOTE Experiment Setting

We over-sampled the minority class in the training data set to control for the imbalanced data distribution by using SMOTE. We ran linear SVC in sets of experiments and we only reported the four best AUC and F-measure for each of the feature group. Table 8 shows the range of AUC between 98.31 and 99.18, whilst F-measure was between 94.57 and 97.14 in the training data set. Additionally, the ranges of AUC and F-measure for testing data set were between 92.36 to 92.68 and 45.83 to 47.98, respectively. In summary, all features in the training data set have good performance. However, only the features character TF-IDF and word embedding showed the most significant performance in the testing data set. After over-sampling the minority class in the training data set, the models performed better compared with the training data set in the default experiment setting. This resulted showed that the linear SVC model generalised well after both class distributions were balanced. However, we used the original unseen data without applying SMOTE in the testing data set because the aforementioned data are not valid for testing purposes (i.e. testing data should only contain real data).

Table 8: Performance evaluation of the linear SVC for the training and testing data sets (%) according to different metrics (A: accuracy; P: precision; R: recall; F_1: F-measure and AUC: area under the ROC curve) under the SMOTE experiment setting

Feature	Training Dataset Score (%)					Testing Dataset Score (%)				
	A	P	R	F_1	AUC	A	P	R	F_1	AUC
One feature: B	94.48	93.13	96.05	94.57	98.31	91.38	33.15	74.23	45.83	92.68
Two features: B + C	94.82	93.45	96.39	94.90	98.46	91.57	33.50	72.80	**45.88**	**92.68**
Three features: A + B + C	96.91	95.19	98.82	96.97	99.13	92.36	35.76	69.91	47.98	92.43
Four features: A + B + C + D	97.01	95.16	98.98	**97.14**	**99.18**	92.50	36.93	79.48	47.11	92.36

4.2.3 Result of Linear SVC with Hyperparameter Optimisation Experiment Setting

We ran 15 set experiments of linear SVC with hyperparameter optimisation to determine the most significant feature group to improve the performance of algorithms. Setting of hyperparameters is shown in Table 3. Moreover, different combinations of linear SVC hyperparameters were ran iteratively on different features to determine the most significant combination of hyperparameters. Table 9 compares the performance evaluation of the training and testing data sets based on five metrics. In the training data set, AUC has varied scores between 97.54 and 98.92, whilst F-measure has varied scores between 50.72 and 63.16. AUC percentages in the testing data varied between 92.37 and 93.17. F-measure varied between 42.60 and 46.72. The best AUC is 98.92 in the training data set for all features with F-measure at 63.16. However, in testing data set, 92.51% of AUC based on character TF-IDF and term list features with F-measure around 44.72%.

Table 9: Performance evaluation of the linear SVC for the training and testing data sets (%) according to different metrics (A: accuracy; P: precision; R: recall; F_1: F-measure and AUC: area under the ROC curve) under the hyperparameter optimisation experiment setting

Feature	Best Hyperparameters	Training Dataset Score (%)					Testing Dataset Score (%)				
		A	P	R	F_1	AUC	A	P	R	F_1	AUC
One feature: B	C: 1, class_weight: balanced, loss: squared_hinge, max_iter: 120000, multi_class: ovr	91.27	34.62	94.81	50.72	97.54	89.77	29.40	77.30	42.60	93.17
Two features: B + D	C: 1, class_weight: balanced, loss: squared_hinge, max_iter: 120000, multi_class: ovr	92.43	38.11	95.87	55.54	98.03	91.00	31.70	75.86	**44.72**	**92.51**
Three features: A + B + D	C: 1, class_weight: balanced, loss: squared_hinge, max_iter: 120000, multi_class: ovr	94.41	45.86	98.96	62.67	98.90	91.90	34.45	72.07	46.62	92.44
Four features: A + B + C + D	C: 1, class_weight: balanced, loss: squared_hinge, max_iter: 120000, multi_class: ovr	94.53	46.38	98.42	**63.16**	**98.92**	91.95	34.61	71.90	46.72	92.37

4.2.4 Result of Linear SVC with Hyperparameter Optimisation and SMOTE Experiment Setting

Table 10 shows the best result of the model performance when we ran 15 sets of experiments in the hyperparameter optimisation and SMOTE setting. In the training data set, the best model performed when implementing all features with AUC 99.24 and F-measure 97.38. In the testing data set, the best model performed when implementing word TF-IDF, character TF-IDF and word embedding with AUC of 92.43 and F-measure of 47.32.

Table 10: Performance evaluation of the linear SVC for the training and testing data sets (%) according to different metrics (A: accuracy; P: precision; R: recall; F$_1$: F-measure and AUC: area under the ROC curve) under the hyperparameter optimisation and SMOTE experiment setting

Feature	**Best Hyperparameters**	**Training Dataset Score (%)**					**Testing Dataset Score (%)**				
		A	**P**	**R**	**F$_1$**	**AUC**	**A**	**P**	**R**	**F$_1$**	**AUC**
One feature: A	C: 1, class_weight: balanced, loss: hinge, max_iter: 120000, multi_class: crammer_singer	86.84	82.18	94.09	87.73	93.12	78.65	15.72	76.76	26.09	85.38
Two features: A + B	C: 1, class_weight: balanced, loss: squared_hinge, max_iter: 120000, multi_class: ovr	96.84	95.20	98.64	96.89	99.13	92.55	36.50	69.91	47.96	92.34
Three features: A + B + C	C: 1, class_weight: balanced, loss: squared_hinge, max_iter: 120000, multi_class: ovr	96.91	95.20	98.82	96.97	99.12	92.36	35.76	70.00	**47.32**	**92.43**
Four features: A + B + C + D	C: 1, class_weight: balanced, loss: squared_hinge, max_iter:120000, multi_class: ovr	97.33	95.68	99.13	**97.38**	**99.24**	92.71	36.93	68.47	47.98	91.87

5.0 DISCUSSION AND CONCLUSION

Numerous researchers globally have been studying information retrieval and linguistics computerisation. As responsible people, we participate to contribute in this endeavour. Thus, this study proposed a systematic framework of cyberbullying detection model in machine learning by involving hyperparameter optimisation and SMOTE as over-sampling technique. Amongst social networks, we used the corpus from ASKfm because there have been cases of suicides caused by cyberbullying, as reported in [87]. On the basis of state-of-the-art cyberbullying detection in machine learning and NLP, we extracted content-based features, namely, word TF–IDF, character TF–IDF, pre-trained Word2vec word embedding and term list, which consists of profane words, proper nouns, negation words, ‘allness’ terms, diminishers and intensifiers. Given that linear SVC is amongst the good algorithms in terms of performance evaluation in binary classification, we utilised this algorithm to classify messages as cyberbullying or non-cyberbullying in the testing data. A total of 60 sets of experiments were conducted to determine the best predictive cyberbullying model.

As shown in Table 7, the combination of word TF–IDF, character TF–IDF and word embedding provides the highest F-measure and AUC in the training and testing data sets. This result may cause no intervention of hyperparameter optimisation and over-sampling techniques. Nonetheless, the AUC score remained lower compared with the other three experimental settings. Table 8 shows that all features contribute the highest F-measure and AUC when over-sampling the data set. After applying the over-sampled technique, features of character TF–IDF and word embedding provides better AUC score. However, F-measure is low because we tested the model on an unseen data set.

When we ran all proposed features under the hyperparameter optimisation setting without over-sampling technique, such as that shown in Table 9, the trained model was able to contribute the highest AUC in the training data set. That is, its performance is slightly better than those of all proposed features with hyperparameter optimisation and over-sampling the data set. Consequently, all proposed features contributed the most when all of them are combined compared if we used a single or double feature when tuning hyperparameters and implementing SMOTE for over-sampling technique to train the model. When the model was tested in the testing data set, all features can still contribute significant percentages of AUC and F-measure, as shown in Table 10. Overall, machine learning working in binary classification should be in balanced data set with the significant hyperparameters of the classifier and should have significant extracted features.

Similar to our previous study, we only extracted word n-grams and conducted experiments in different settings, which consist of hyperparameter optimisation, chi-square feature selection and over-sampled using the SMOTE technique [88]. Overall, performance evaluation of the training and testing data sets in this study is better than the previous one, even though we did not implement the feature selection technique. The types of features used in cyberbullying detection could have an impact on the effectiveness of the built model.

Amongst the contributions of this study is to provide evidence that social media and machine learning technique can be valuable, particularly in text classification, by using the appropriate features. Text classification skills are essential for data scientists because these techniques are at the core of every data analysis. Moreover, the proposed framework could be used in other cases, such as in medical prediction, romance scam deception and spam and fake news detections. The outcomes of this study could be used to create customised and new data models. It also aids the mining industry for an improved understanding of mining processes, apart from supporting effective decision-making.

6.0 FUTURE STUDY

We comprehend that a paradigm shift is necessary in future research on text classification machine learning, specifically in cyberbullying detection. As another fascinating direction of future research, we will conduct experiments on different algorithms, such as naïve Bayes, decision tree, logistic regression and random forest. We will also explore other features, such as sentiment-based features (i.e. polarity) and user-based features (i.e. age, gender, race). Additionally, deep learning could contribute in binary classification, which we could explore in our future research.

7.0 LIMITATION

The current study has a few limitations. Firstly, we were unable to conduct comprehensive analysis because the data set have no relation with other metadata user-based features (i.e. age, gender, race) and network-based features (i.e. user-interaction, ego network, time online) [39]. Secondly, we also failed to present the literature on cyberbullying detection modelling in machine learning techniques that does not involve languages other than English owing to our own linguistic limitation. Thirdly, we were unable to provide a comprehensive overview of the study because we are still working on feature engineering involving sentiment-based features and another content-based features, such as n-grams and part-of-speech tagging (PoS-tagging).

8.0 ACKNOWLEDGEMENT

The authors would like to extend their gratitude to Cynthia Van Hee for sharing the AMiCA Bullying Cyber Dataset. We will forever grateful for your cooperation.

REFERENCES

[1] B. M. Leiner et al., "Brief History of the Internet*", Internet Society,* vol. 36. 1997, pp. 1–19.

[2] S. Edosomwan, S. K. Prakasan, D. Kouame, J. Watson, and T. Seymour, "The History of Social Media and its Impact on Business" *Management*, vol. 16, no. 3, 2011, pp. 79–91.

[3] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media" *Bus. Horiz.*, vol. 53, no. 1, 2010, pp. 59–68.

[4] G. P. Kumar and M. Vasimalaraija, "Benefits of Using Social Media" *J. Chem. Inf. Model.*, vol. 53, no. 9, 2008, p. 287.

[5] M. Dhingra and R. K. Mudgal, "Historical Evolution of Social Media: An Overview" *International Conference on Advances in Engineering Science Management & Technology (ICAESMT),* Uttaranchal University, Dedradun, India, 2019, pp.1-8.

[6] R. Faizi, A. El Afia, and R. Chiheb, "Exploring the Potential Benefits of Using Social Media in Education" *Int. J. Eng. Pedagog.*, vol. 3, no. 4, 2013, pp. 50-53.

[7] B. S. Nandhini and J. I. Sheeba, "Online Social Network Bullying Detection Using Intelligence Techniques" *Procedia - Procedia Comput. Sci.*, vol. 45, 2015, pp. 485–492.

[8] N. M. Zainudin, K. H. Zainal, N. A. Hasbullah, N. A. Wahab, and S. Ramli, "A review on cyberbullying in Malaysia from digital forensic perspective" *ICICTM 2016 - Proc. 1st Int. Conf. Inf. Commun. Technol.,* no. May, Kuala Lumpur, 3 April 2017, 2017, pp. 246–250.

[9] S. Nadali, M. A. A. Murad, N. M. Sharef, A. Mustapha, and S. Shojaee, "A review of cyberbullying detection: An overview" *Int. Conf. Intell. Syst. Des. Appl. ISDA,* Selangor, 13 October 2013, vol. 13, pp. 325–330.

[10] E. Aboujaoude, M. W. Savage, V. Starcevic, and W. O. Salame, "Cyberbullying: Review of an old problem gone viral" *J. Adolesc. Heal.*, vol. 57, no. 1, 2015, pp. 10–18.

[11] N. Tarmizi, S. Saee, and D. H. A. Ibrahim, "Detecting the usage of vulgar words in cyberbully activities from Twitter," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 10, no. 3, 2020, pp. 1117–1122.

[12] H. Vandebosch and K. Van Cleemput, "Defining cyberbullying: A qualitative research into the perceptions of youngsters," *Cyberpsychology Behav.*, vol. 11, no. 4, 2008, pp. 499–503.

[13] N. Willard, "Educator ' s Guide to Cyberbullying , Cyberthreats & Sexting," *Center for Safe and Responsible Use of the Internet*, 2007.

[14] G. S. O' Keeffe and K. Clarke-Pearson, "The impact of social media on children, adolescents, and families" Official journal of the American Academy of Pediatrics, 2011.

[15] J. M. Xu, K. S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media" *NAACL HLT 2012 - 2012 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, Montreal, 03 June 2012, pp. 656–666.

[16] D. Pozza, A. Di Pietro, S. Morel, and E. Psaila, "Cyberbullying among young people.", 2016, pp. 1-194.

[17] R. Slonje and P. K. Smith, "Cyberbullying : Another main type of bullying ?" *Scand. J. Psychol.*, vol. 49, 2008, pp. 147–154.

[18] P. Vuadens, "Definition and Measurement of Outcome" *Long-Term Eff. Stroke*, vol. 4, no. 2, 2002, pp. 1–12.

[19] A. Saravanaraj, J. I. Sheebaassistant, S. Pradeep, and D. Dean, "Automatic Detection of Cyberbullying From Twitter", *IRACST -International J. Comput. Sci. Inf. Technol. Secur.*, vol. 6, no. 2016, 2249–9555.

[20] J. W. Patchin and S. Hinduja, "Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying", *Youth Violence Juv. Justice*, vol. 4, no. 2, 2006, pp. 148–169.

[21] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: Its nature and impact in secondary school pupils", *J. Child Psychol. Psychiatry Allied Discip.*, vol. 49, no. 2008, 376–385.

[22] H. Vandebosch and K. van Cleemput, "Cyberbullying among youngsters: Profiles of bullies and victims" *New Media Soc.*, vol. 11, no. 8, 2009, pp. 1349–1371.

[23] N. E. Willard, "The Authority and Responsibility of School Officials in Responding to Cyberbullying," *J. Adolesc. Heal.*, vol. 41, no. 6, 2007, pp. 64–65.

[24] J. J. Dooley, J. Pyzalski, and D. Cross, "Cyberbullying versus face-to-face bullying: A theoretical and conceptual review" *J. Psychol.*, vol. 217, no. 4, 2009, pp. 182–188.

[25] A. A. H. Shafie, A. A. Anuar, N. Che Rozudi, W. A. Z. Wan Kamaruddin, and M. Mohamd, "Mangsa buli dan kesan buli", *J. Islam. Soc. Sci. Humanit.*, vol. 11, 2017, pp. 109–124.

[26] M. Teicher, M. Hospital, J. A. Samson, M. Hospital, Y. Sheu, and A. Polcari, "Hurtful Words: Exposure to Peer Verbal Aggression is Associated with Elevated Psychiatric Symptom Scores and Corpus Callosum Abnormalities", *Am J Psychiatry*, vol. 167, no. 12, 2010, pp. 1464-1471.

[27] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network", *Comput. Human Behav.*, vol. 63, 2016, pp. 433–443.

[28] C. Van Hee et al., "Automatic detection of cyberbullying in social media text", *PLoS One*, vol. 13, no. 10, 2018, pp. 1–21.

[29] B. A. Talpur and D. O'Sullivan, "Cyberbullying severity detection: A machine learning approach", *PLoS One,* vol. 15, no. 10 October, 2020, pp. 1–19.

[30] R. Zhao, A. Zhou, and K. Mao, "Automatic Detection Of cyberbullying on Social Networks Based on Bullying Features", *Proc. 17th Int. Conf. Distrib. Comput. Netw. - ICDCN '16*, Singapore, 4-7 January 2016, pp. 1–6.

[31] T. O. Ayodele, "Types of Machine Learning Algorithms", in *New Advances in Machine Learning*, Y. Zhang, Ed. Shanghai: InTech, 2010, pp. 3–31.

[32] Wikipedia, "Ask.fm", *Encyclopedia of Social Network Analysis and Mining*, 2021. https://en.wikipedia.org/wiki/Ask.fm

[33] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", *J. Artif. Intell. Res. 16*, vol. 16, 2002, pp. 321–357.

[34] M. A. Al-Garadi *et al.*, "Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges," *IEEE Access*, vol. 7, 2019, pp. 70701–70718.

[35] J. Kerstens and S. Veenstra, "Cyber bullying in the Netherlands: A criminological perspective", *Int. J. Cyber Criminol.*, vol. 9, no. 2, 2016, pp. 144–161.

[36] T. Alsubait and D. Alfageh, "Comparison of Machine Learning Techniques for Cyberbullying Detection on YouTube Arabic Comments," *IJCSNS International Journal of Computer Science and Network Security*, vol. 21, no. 1, 2021, pp. 1–5.

[37] J. W. Patchin and S. Hinduja, "*Words Wound*," Minneapolis, Free Spirit Publishing Inc, 2014.

[38] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," *Proc. 9th Int. Conf. Web Soc. Media, ICWSM 2015*, 2015, pp. 61–70.

[39] S. Salawu, Y. He, and J. Lumsden, "Approaches to Automated Detection of Cyberbullying: A Survey", *IEEE Trans. Affect. Comput.*, 2017, pp. 1–25.

[40] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying", in *The Social Mobile Web 2011*, 2011, pp. 11–17.

[41] V. Nahar, X. Li, and C. Pang, "An Effective Approach for Cyberbullying Detection", *Commun. Inf. Sci. Manag. Eng.*, vol. 3, no. 5, 2013, pp. 238–247.

[42] K. Reynolds, A. Kontostathis, and L. Edwards, "Using Machine Learning to Detect Cyberbullying", *Proc. Tenth Int. Conf. Mach. Learn. Appl.*, 2011, pp. 1–4.

[43] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of Harassment on Web 2.0", *Proc. Content Anal. WEB .*, vol. 2, 2009, pp. 1–7.

[44] A. Mahmud, K. Z. Ahmed, and M. Khan, "Detecting flames and insults in text", *Proc. 6th Int. Conf. Nat. Lang. Process*, 2008, pp. 1–10.

[45] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. De Jong, "Improving cyberbullying detection with user context", *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7814 LNCS, 2013, pp. 693–696.

[46] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean Birds: Detecting Aggression and Bullying on Twitter", *Proc. 2017 ACM web Sci. Conf.*, New York, 25-28 June 2017, pp. 13–22.

[47] K. R. Talpur, S. S. Yuhaniz, N. N. B. A. Sjarif, and B. Ali, "Cyberbullying detection in roman urdu language using lexicon based approach", *J. Crit. Rev.*, vol. 7, no. 16, 2020, pp. 834–848.

[48] H. Rosa, D. Matos, R. Ribeiro, L. Coheur, and J. P. Carvalho, "A 'Deeper' Look at Detecting Cyberbullying in Social Networks", *Proc. Int. Jt. Conf. Neural Networks*, Rio de Janeiro, 15 October 2018, pp. 1-8.

[49] K. Stoitsas, "The use of word embeddings for cyberbullying detection in social media", Netherland, July 2018, pp. 1–2.

[50] M. Dadvar, D. Trieschnigg, and F. De Jong, "Expert knowledge for automatic detection of bullies in social networks", *Conf. 25th Benelux Conf. Artif. Intell.*, 2013, pp. 1–7.

[51] V. S. Chavan and S. S. Shylaja, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," *2015 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2015*, India, 10-13 August 2015, pp. 2354–2358.

[52] A. Perera and P. Fernando, "Accurate Cyberbullying Detection and Prevention on Social Media," *Procedia Comput. Sci.*, vol. 181, 2021, pp. 605–611.

[53] R. Shah, S. Aparajit, R. Chopdekar, and R. Patil, "Machine Learning based Approach for Detection of Cyberbullying Tweets," *Int. J. Comput. Appl.*, vol. 175, no. 37, 2020, pp. 51–56.

[54] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, "Social media cyberbullying detection using machine learning", *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 5, 2019, pp. 703–707.

[55] X. Zhang *et al.*, "Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network" *2016 15th IEEE Int. Conf. Mach. Learn. Appl.*, Anaheim USA, 18-20 December 2016, pp. 740–745.

[56] M. A. Al-Ajlan and M. Ykhlef, "Deep learning algorithm for cyberbullying detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 9, 2018, pp. 199–205.

[57] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," *Multimed. Syst.*, 2020, pp. 1-13.

[58] P. Stenetorp, S. Pyysalo, and G. Topi, "BRAT : a Web-based Tool for NLP-Assisted Text Annotation", *EACL '12 Proc. Demonstr. 13th Conf. Eur. Chapter Assoc. Comput. Linguist.*, France, April 2012, pp. 102–107.

[59] Group Research Computational Linguistics, "Bratreader." Antwerp, Belgium, *https://github.com/clips/bratreader/*, 2019.

[60] M. Z. H. Jesmeen *et al.*, "A survey on cleaning dirty data using machine learning paradigm for big data analytics", *Indones. J. Electr. Eng. Comput. Sci.*, vol. 10, no. 3, 2018, pp. 1234–1243.

[61] P. Jeatrakul, K. W. Wong, and C. C. Fung, "Data cleaning for classification using misclassification analysis", *J. Adv. Comput. Intell. Intell. Informatics*, vol. 14, no. 3, 2010, pp. 297–302.

[62] Y. J. Foong and M. Oussalah, "Cyberbullying System Detection and Analysis", *Eur. Intell. Secur. Informatics Conf. Cyberbullying*, 2017, pp. 40–46.

[63] S. Pericherla and E. Ilavarasan, "Performance analysis of Word Embeddings for Cyberbullying Detection", *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1085, no. 1, 2021, pp. 1–11.

[64] D. Zhang, H. Xu, Z. Su, and Y. Xu, “Chinese comments sentiment classification based on word2vec and SVMperf”, *Expert Syst. Appl.*, vol. 42, no. 4, 2015, pp. 1857–1863.

[65] G. Jacobs, C. Van Hee, and V. Hoste, “Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text?”, *Nat. Lang. Eng.*, 2020, pp. 1–26.

[66] F. Sommar and M. Wielondek, “Combining Lexicon- and Learning-based Approaches for Improved Performance and Convenience in Sentiment Classification”, *KTH Royal Institute of Technology,* May 2015, pp. 1-50.

[67] F. Strohm, “The Impact of Intensifiers , Diminishers and Negations on Emotion Expressions", University of Stuttgart, 03 August 2017, pp. 1-73.

[68] C. E. Osgood and E. G. Walker, “Motivation and language behavior: A content analysis of suicide notes”, *J. Abnorm. Soc. Psychol.*, vol. 59, no. 1, 1959, pp. 58–67.

[69] T. Tytko and M. C. Augstkalns, “How well do we know ourselves ? Identifying suicide markers in online communication : A case study of a graduate student ’ s writing”, University of Maryland College Park, 2020. pp. 45–62.

[70] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, “A lexicon-based approach for hate speech detection”, *Int. J. Multimed. Ubiquitous Eng.*, vol. 10, no. 4, 2015, pp. 215–2301.

[71] W. M. Czarnecki, S. Podlewska, and A. J. Bojarski, “Robust optimization of SVM hyperparameters in the classification of bioactive compounds”, *J. Cheminform.*, vol. 7, no. 1, 2015, pp. 1–15.

[72] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, and B. Thirion, “Scikit-learn: Machine Learning in Python”, *J. ofMachine Learn. Res.*, vol. 12, no. 9, 2011, pp. 2825–2830.

[73] M. Sintaha, S. Satter, N. Zawad, C. Swarnaker, and A. Hassan, “Cyberbullying Detection Using Sentiment Analysis in Social,” BRAC University, 2016.

[74] R. Blagus and L. Lusa, “Open Access SMOTE for high-dimensional class-imbalanced data”, *BMC Bioinformatics*, vol. 14, no. 106, 2013, pp. 1–16.

[75] N. Qazi and K. Raza, “Effect of feature selection, Synthetic Minority Over-sampling (SMOTE) and under-sampling on class imbalance classification,” *2012 14th Int. Conf. Model. Simul.*, Cambridge UK, 28-30 March 2012, pp. 145–150.

[76] M. Feurer and F. Hutter, *Hyperparameter Optimization*, Automated Machine Learning. The Springer Series on Challenges in Machine Learning. Springer, Cham, 2019, pp. 3-33.

[77] E. K. Hashi and Md. Shahid Uz Zaman, “Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction”, *J. Appl. Sci. Process Eng.*, vol. 7, no. 2, 2020, pp. 631–647.

[78] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, “Systematic ensemble model selection approach for educational data mining”, *Knowledge-Based Syst.*, vol. 200, 2020, pp. 1-16.

[79] J. W. Tukey, “The Future of Data Analysis”, *Ann. Math. Stat.*, vol. 33, no. 1, 1962, pp. 1–67

[80] B. Tessem, S. Bjørnestad, W. Chen, and L. Nyre, “Word cloud visualisation of locative information”, *J. Locat. Based Serv.*, vol. 9, no. 2015, pp. 254–272.

[81] K. H. Zou, A. J. O’Malley, and L. Mauri, “Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models”, *Circulation*, vol. 115, no. 5. 2007, pp. 654–657.

[82] T. Fawcett, “An introduction to ROC analysis”, *Pattern Recognit. Lett.*, vol. 27, no. 8, 2006, pp. 861–874.

[83] Y. Liu, Y. Zhou, S. Wen, and C. Tang, “A Strategy on Selecting Performance Metrics for Classifier Evaluation”, *Int. J. Mob. Comput. Multimed. Commun.*, vol. 6, no. 4, 2014, pp. 20–35.

[84] S. Visa, B. Ramsay, A. Ralescu, and E. Van Der Knaap, “Confusion matrix-based feature selection,” *Proceedings of The 22nd Midwest Artificial Intelligence and Cognitive Science Conference 2011*, Cincinnati, Ohio, USA, 16-17 April, 2011, vol. 710, pp. 120–127.

[85] D. M. W. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation,” *Int. J. Mach. Learn. Technol.*, vol. 2, no. 1, 2011, pp. 37–63.

[86] H. M and S. M.N, “A Review on Evaluation Metrics for Data Classification Evaluations,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, 2015, pp. 01–11.

[87] A. A. Santos, “Ask.fm Is the New Way for Teens to Cyberbully Each Other to Death”, https://www.theatlantic.com/international/archive/2013/08/twitter-bullying-over-ask-new-way-people-be-awful/312421/, 2013.

[88] W. N. H. W. Ali, M. Mohd and F. Fauzi, "Cyberbullying Predictive Model: Implementation of Machine Learning Approach," *2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP)*, Kuala Lumpur, Malaysia, 15-16 June, 2021, pp. 65-69.