Co-funded by the
Erasmus+ Programme
of the European Union

The Comparative analysis of different predictive analytics models in predicting cyberbullying

MASTER THESIS
to obtain the Erasmus Mundus Joint Master Degree in Digital Communication Leadership (DCLead)

of

Faculty of Social Sciences

Paris Lodron University of Salzburg, Austria

Technical Faculty of IT and Design

Aalborg University in Copenhagen, Denmark

Submitted by
[MUHAMMAD WAJAHAT, AMJAD]
s1093322

Primary Supervisor: Reza Tadayoni
Secondary Supervisor: Badr Hanan

Department of Communication Studies
Salzburg, [30-0-2024]

1. Table of Contents

2. Table of figures

3. Executive Summary

The emergence of social media, especially X, has led to misunderstandings about freedom of speech, resulting in issues like cyberbullying. This study investigates cyberbullying's effects by comparing machine learning algorithms based on accuracy, precision, recall, and F1 score. It also assesses how sentiment analysis and Psychosocial Safety Climate (PSC) principles can improve model efficiency.

Utilizing PSC theory, which promotes prosocial behavior to prevent bullying, the study merges technological solutions with human behavior insights for online safety. Analyzing 2,000 tweets using TextBlob for sentiment analysis, features are extracted through TF-IDF vectorization, and models are trained using Multinomial Naive Bayes, Random Forests, XGBoost, and Support Vector Machines (SVM). These models are evaluated using 5-fold cross-validation, with hyperparameters fine-tuned by GridSearchCV.

Results show that Random Forest and XGBoost achieved the highest scores, 0.761 and 0.740, respectively. Multinomial Naive Bayes demonstrated exceptional computational efficiency, making it suitable for real time applications. Sentiment analysis improved detection by emphasizing emotional context, and PSC principles enhanced model effectiveness by incorporating features like "number_negative_words" and "number_positive_words."

The research underscores the combination of machine learning and psychosocial theory in detecting cyberbullying. It recommends choosing models based on application needs: Random Forest for a balance between performance and interpretability, XGBoost for high accuracy, and Multinomial Naive Bayes for efficiency. Future research should expand datasets, address privacy concerns, and incorporate features like social network analysis to enhance practicality and improve online safety by involving administrators and moderators.

Keywords: Cyberbullying, Predictive Analytics, Sentiment Analysis, Multinomial Naive Bayes, Random Forests, XGBoost, Support Vector Machines, PSC

4. Introduction

4. 1 Background and motivation

With the progress of technology, the usage and utilization of social media is expanding every day. According to Global social media statistics, as of 2024 (Larson, 2024) globally, 5.16 billion individuals are active social media users which is around 59.3 percent of world population and is projected to increase to 5.84 billion by 2027 as shown in the figure 1 below. Networking sites like Facebook, Instagram, and X are being used to communicate in real time worldwide.

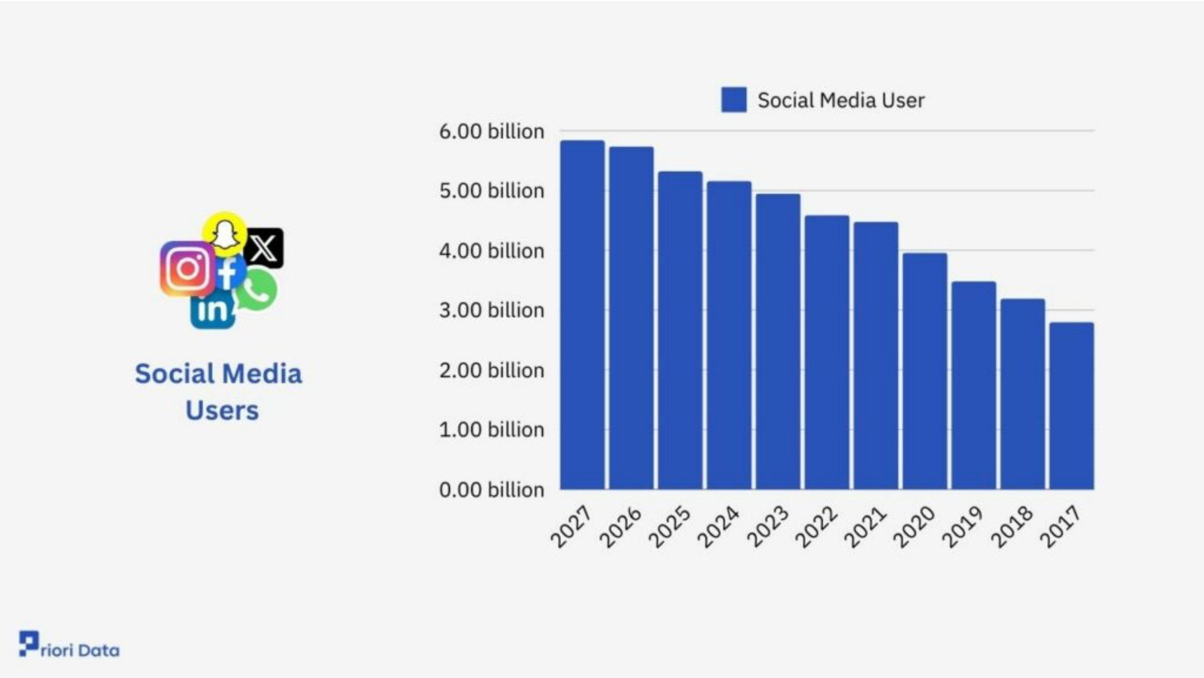


Figure 1 Social Media Users: 2017 to 2027 (Larson, 2024)

However, the tech revolution has a negative aspect, like cybercrimes most notably cyberbullying. Cyberbullying is defined as repeatedly and intentionally mistreating, harassing or making fun of another person online while using digital devices such as cell phones or computers. According to (Affairs (ASPA), 2019), cyberbullying includes sending, sharing or posting harmful, mean or false content about someone else or sharing private or personal information about someone with the intent of causing humiliation or embarrassment. Cyberbullying is a growing problem, according to the most recent study conducted by the Cyberbullying Research Center, approximately 55% of students report having been cyberbullied once during their lifetime and about 27% saying it happened within the previous 30 days (J. W. Patchin, 2024).

Despite being a fairly broad term, there is still disagreement about what constitutes cyberbullying. It is challenging to define it precisely and consistently due to a variety of reasons. Online environments are characterised by a number of features that are exclusive to bullying, such as its recurrent nature, power imbalances, the prevalence of aggressive behaviors, anonymity, and publicity (Perera & Fernando, 2021). To put it another way, as Figure 2 illustrates, a cyberbully may post offensive, mean, or threatening content that records the target's attitudes, actions, and choices. The following categories to classify cyberbullying broadly are discussed upon various platforms such as Social Media Victims Law Center (*Types of Cyberbullying - Examples of Bullying Online*, n.d.):

1. **Flaming:** This is the term for online conflict that arises when someone posts hurtful, hateful, or disparaging remarks on their blog, website, or social media pages.
2. **Impersonation:** This occurs when bullies create false accounts to mimic the victim and then post hurtful or offensive comments while posing as the victim. Inappropriate comments can be posted by bullies using their account, which they have gained access to through password theft or other means.
3. **Exclusion:** This is organizing gatherings or organizations and purposefully leaving someone out. It may also entail purposefully leaving the victim out of or neglecting them in an online discussion, as well as failing to tag them in a picture.
4. **Harassment:** refers to the practice of frequently sending threatening, abusive, or offensive messages to a person or group via social media using direct messages (DMs), instant messages (IMs), etc.
5. **Cyberstalking:** Cyberstalking is the practice of continuously threatening, intimidating, or harassing someone by using technology to monitor their targets. Cyberstalkers typically try to meet their targets, and occasionally they groom teenagers with the intention of having sex with them.

6. **Outing:** This is when the victim is tricked by the bully into disclosing private information that the cyberbully then posts online. Alternatively, the victim could be threatened with blackmail if the cyberbully obtains the information and threatens to share it with others online.
7. **Denigration/Gossip:** this type of behavior entails spreading rumors or gossip about another person online, usually with the goal of destroying the target's relationships and reputation.
8. **Cyber threats:** Cyberbullies who suggest or threaten to use violence against others are posing a threat to others online.

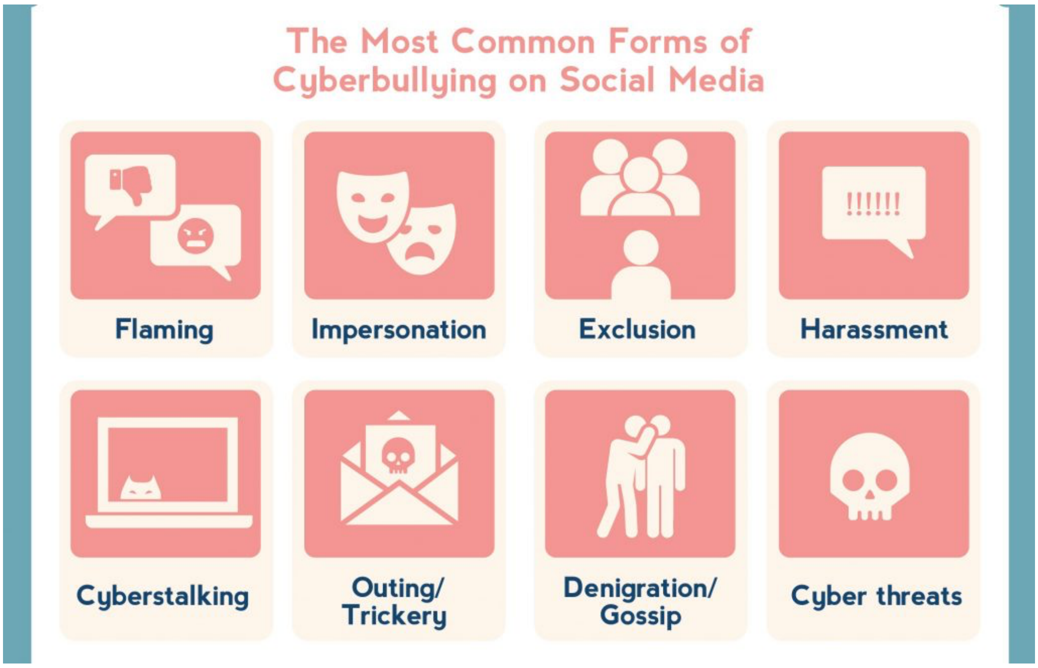


Figure 2 Types of Cyberbullying (Cyberbullying On Social Media, 2019)

Compared to traditional bullying, cyberbullying may be more damaging to victims due to a few key distinctions. Seeking refuge is simpler in cases of physical bullying than it is in cases of cyberbullying, where the dissemination of online content cannot be stopped. Put differently, cyberbullies might post and disseminate the victim's embarrassing photos once, but as more people like and share the content, the victimization happens again without the bullying act. According to Patchin "Adolescent girls are more likely to have experienced cyberbullying in their lifetimes (59.2% vs.

49.5%). This difference is not as dramatic when reviewing experiences over the previous 30 days, where rates are more similar (24.2% of boys and 28.6% of girls have been cyberbullied recently), though differences in lifetime and 30-day rates are both statistically significant (*p* < .001). " (J. W. Patchin, 2024).

Many victims of bullying experience traumatic stress, anxiety, disconnection from their communities and schools, and loss of interest in what they once enjoyed. Bullying has the potential to negatively impact a victim's mental health and make them disinterested in their hobbies. The eating and sleeping habits of teenagers are also depicted in Figure 3, and these behaviors may be influenced by depressive episodes, low self esteem, and thoughts of suicide. According to data from Centers for Disease Control and Prevention, 14.9% of teenagers have experienced cyberbullying, and 13.6% have attempted suicide with serious consequences (Schonfeld et al., 2023).

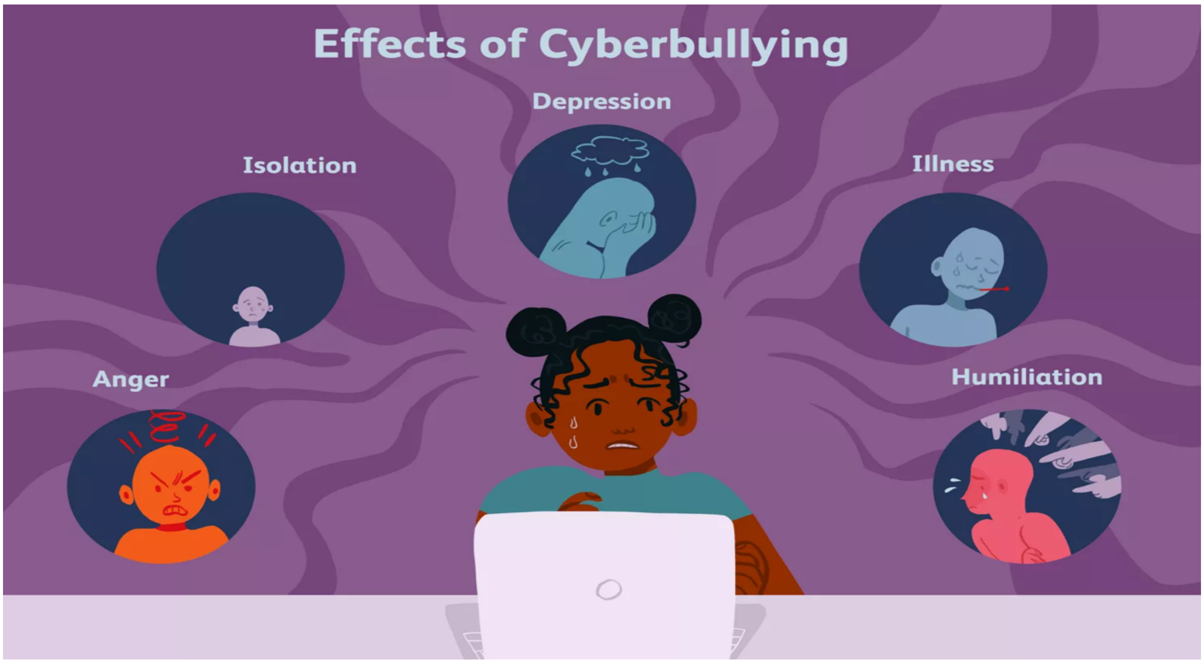


Figure 3 Effects of Cyberbullying(Learn to Recognize the Real-Life Effects of Cyberbullying on Children, n.d.)

Given the consequences, it is essential to exercise caution when it comes to content found online. When 2020 X samples were being examined, an average of 194,444 tweets in different languages were sent every minute (Ates et al., 2021). With an average of 279,999,360 tweets posted in a 24 hour sample, it would be nearly impossible to physically review such a large volume of data in multiple languages (Ates et al., 2021). As a result, the majority of current research focuses on machine learning based categorization after text normalization via NLP techniques. Machine learning may be used to identify language patterns that bullies and their victims use, and it may be possible to build a system that automatically recognizes cyberbullying content and produces statistical analysis.

4. 2 Research Questions

Given the rapid growth and pervasiveness of social media, there is an urgent need for effective detection and intervention strategies due to the rising prevalence of cyberbullying. Cyberbullying has serious psychological and social repercussions that include anxiety, depression, and even suicidal thoughts and actions, underscoring its crucial social significance. Traditional detection methods face significant challenges due to the multifaceted nature of cyberbullying, which includes exclusion, harassment, flaming, and impersonation, as discussed in the background section. Because of this complexity, a thorough, nuanced strategy utilizing cutting-edge machine learning and predictive analytics models is required.

In light of this, the following research questions has been selected in an attempt to investigate how well different machine learning algorithms detect cyberbullying. They also look into how adding sentiment analysis can improve the predictive ability of these models and how combining Psychosocial Safety Climate (PSC) principles can lead to a more thorough understanding and identification of cyberbullying incidents. In addition to being technologically and socially significant, these investigations aim to reduce the negative impacts of cyberbullying by creating dependable, expandable, and ethically sound detection systems.

How do different machine learning algorithms (e.g., Logistic Regression, Random Forest, SVM) compare in terms of accuracy, precision, recall, and F1 score for cyberbullying detection?

a. How does the inclusion of sentiment analysis features influence the effectiveness of predictive models in identifying cyberbullying?

b. How can principles of Psychosocial Safety Climate (PSC) be integrated into predictive analytics models to enhance cyberbullying detection?

4. 3 Research Objective

The main goal of this research is to perform a thorough comparative analysis of different predictive analytics models, such as Logistic Regression, Random Forest, Support Vector Machine, and others, in order to assess their efficacy in identifying instances of cyberbullying on social media platforms. This study seeks to improve the accuracy and reliability of models by combining advanced natural language processing techniques with Psychosocial Safety Climate (PSC) principles. In addition, the research aims to determine how different methods of preparing text and linguistic characteristics affect the performance of the model. It also aims to tackle practical challenges in implementing the model in real world scenarios and examine the ethical considerations of using such models. The ultimate objective is to create a strong framework for detecting cyberbullying that can be seamlessly integrated into preventive and intervention strategies, thus reducing the detrimental impact of cyberbullying.

4. 4 Structure

This research paper, titled "The Comparative Analysis of Different Predictive Analytics Models in Predicting Cyberbullying," systematically examines multiple aspects. The study commences with an Introduction that presents the Background and Motivation, emphasizing the social significance of cyberbullying and the necessity for sophisticated detection techniques. The investigation is guided by the Research Questions and

Research Objectives. The Literature Review analyzes prior research on the detection of cyberbullying, with a specific emphasis on the utilization of natural language processing (NLP) and machine learning methods. The Theory section presents an analysis of the Psychosocial Safety Climate (PSC) theory, which provides a structured approach to comprehending the psychological and social aspects of cyberbullying. The Methodology provides a comprehensive overview of the research design, encompassing data collection, preprocessing, sentiment analysis, and the utilization of predictive analytics models. The Results section assesses the efficacy of various machine learning algorithms and the influence of sentiment features and PSC principles. The Discussion section of the study provides an interpretation of the findings, examines their implications for the detection of cyberbullying, and proposes potential areas for future research. Ultimately, the paper concludes by succinctly summarizing the valuable understandings gained and offering well founded recommendations. The study is supported by a comprehensive Bibliography and additional materials in the Appendices.

5. Literature Review

In an effort to limit or manage cyberbullying on social media platforms, researchers have been focusing a great deal of effort on cyberbully detection for a number of years. Cyberbullying is concerning because the victims are unable to handle the psychological toll that harsh, threatening, demeaning, and violent messages take. The phenomenon of cyberbullying needs to be investigated in terms of detection, prevention, and mitigation in order to lessen its negative effects. Many international programmes are currently being implemented with the goal of stopping cyberbullies and enhancing internet user safety, particularly for kids (Goldberg & Levy, 2014; Pennington et al., 2014). Numerous studies on what are known as intervention and prevention approaches—a means of preventing cyberbullying—have been published in the literature. These methods have their roots in the educational and psychology domains. Nevertheless, these methods are uncommon worldwide. Furthermore, victims of cyberbullying frequently decline to talk to their parents (LI et al., 2019), teachers (Jiang et al., 2016), or other adults (Hassan Yousef et al., 2014). They spend a lot of time on the internet

(Al-Garadi et al., 2019), frequently ask for help anonymously (*(PDF) Logistic Regression in Data Analysis: An Overview*, n.d.), and post requests for information and support (Jr et al., 2013). Nonetheless, using the Internet to distribute anti-bullying solutions is highly effective. Additionally, patients can use web based approaches whenever and wherever they choose (Chavan & S S, 2015). For example, the Anti-Harassment campaign in France (*(PDF) Cyberbullying in Children and Youth: Implications for Health and Clinical Practice*, n.d.) and the Kiva anti-cyberbullying programme at the University of Turku, Finland (*Cyberbullying: Where Are We Now? A Cross-National Understanding (Printed Edition of the Special Issue Published in Societies). | Request PDF*, n.d.), and an anti-cyberbully initiative by the Belgian government (Goerzig & Ólafsson, 2012). Ideally, these prevention and intervention approaches should: (1) increase awareness of potential cyberbully threats through individualized intensive intervention strategies based on the victims' needs (Mangaonkar et al., 2015; *(PDF) Scalable and Timely Detection of Cyberbullying in Online Social Networks*, n.d.); (2) provide health education and teach emotional self-management skills (*(PDF) XBully: Cyberbullying Detection within a Multi-Modal Context*, n.d.); (3) increase awareness of victims in both reactive measures (e.g., deleting, blocking and ignoring messages), and preventive measures (e.g., increased awareness and security) (Nahar et al., 2014); provide practical strategies and resources that allow victims to cope with experienced stress and negative emotions (Nahar et al., 2014); (4) aim to reduce traditional bullying as well (Mandot, 2018) since victims are often involved in both forms of bullying (*(PDF) Performance Analysis of Boosting Classifiers in Recognizing Activities of Daily Living*, n.d.; Zinovyeva et al., 2020); and (5) include empathy training, Internet labelling and healthy Internet behavior (Ke et al., 2017; Lu et al., 2020). It has proven challenging to stop cyberbullying thus far. The majority of parents and educators rely on kids' awareness of the origins and effects of cyberbullying. Peer-mentoring, according to some parents, is a successful strategy to stop cyberbullying, especially in adolescence when peers have a bigger influence than family and school. To assist the victims, more specialised methods or internet resources must be created (*(PDF) Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language*, n.d.). (Brownlee, 2016), for instance, cautioned against preventing bullying by stating that such programmes only slightly alter student behaviour. In a similar vein, writers in (Pawar, 2018) recommend that when creating their programmes to prevent cyberbullying, schools incorporate the following actions: (1) Identify cyberbullying; (2) Implement robust policies; (3) Educate staff, students, and parents on policy identification; (4) Use internet filtering technologies to guarantee adherence. Previous studies have suggested that social reinforcement could play a major role as a protective factor in reducing the negative consequences of cyberbullying (*(PDF) AUTOMATIC DETECTION OF CYBERBULLYING IN FORMSPRING.ME, MYSPACE AND YOUTUBE SOCIAL NETWORKS*, n.d.; *(PDF) Multilingual Cyberbullying Detection System*, n.d.). They must ask for assistance in order to receive the necessary reinforcement to lessen the associated negative effects of cyberbullying. Nonetheless, some reports indicate that victims of cyberbullying would rather remain silent and are unable to report bullying incidents (Deb, 2016). Few teenagers ever ask their teachers or school advisors for help (Patel, 2017). It is imperative to identify and filter instances of cyberbullying on social media in light of the prevention strategies discussed above. Consequently, the examination of cyberbully detection methods is the focus of this section. According to the literature review, machine learning and natural language processing are the two primary approaches for identifying cyberbullies. These are further discussed in the ensuing subsections.

5. 1 Natural language Process In Cyberbullying Detection

Using natural language processing (NLP) to identify offensive content is one approach in this field. The "levels of language" approach is the most illuminating way to explain what goes on inside a Natural Language Processing system (Louppe, 2014). People utilise these levels to decipher spoken or written languages. The term "levelling" describes the fact that formal models or representations of knowledge pertaining to these levels are the primary source of information used in language processing (Louppe, 2014; Novalita et al., 2019). Furthermore, by utilising linguistic knowledge, language processing apps set themselves apart from data processing systems. The following levels comprise the analysis of natural language processing (Muneer & Fati, 2020a) :

- Phonology level (knowledge of linguistic sounds)
- Morphology level (knowledge of the meaningful components of words)
- Lexical level (deals with the lexical meaning of words and parts of speech analyses)
- Syntactic level (knowledge of the structural relationships between words)
- Semantic level (knowledge of meaning)
- Discourse level (knowledge about linguistic units more extensive than a single utterance)
- Pragmatic level (knowledge of the relationship of meaning to the goals and intentions of the speaker)

For instance, Dinakar et al. (García-Recuero, 2016) employed a common sense knowledge base and related reasoning strategies. Using query terms found in cyberbullying cases, Kontostathis et al. (Waseem & Hovy, 2016) identified cyberbullying content based on Formspring.me data. In order to identify indicators of bullying, Xu et al. (Chengsheng et al., 2017) employ a variety of natural language processing techniques (a novel term referring to online references that could be bullying instances themselves or online references relating to offline bullying cases). Latent Dirichlet Analysis is used to identify subjects/themes, and sentiment analysis features are used to identify bullying roles. The authors of (Chengsheng et al., 2017) hope to lay the groundwork for a number of tasks pertaining to bullying identification and to encourage other researchers to improve these particular methods.
Consequently, the first researchers in NLP cyberbullying detection are Yin et al. (Chatterjee, 2018), Reynolds et al. (Misra & Li, 2020), and Dinakar et al. (García-Recuero, 2016). They studied predictive strength n-grams, part-speech information (e.g., first and second pronouns), and sentiment information based on profanity lexicons for this task (with and without TF-IDF weighting). Similar characteristics were also applied in (Rafiq et al., 2015) to identify instances of cyberbullying and text categories at a finer level.

In conclusion, Term Frequency (TF) (Raza, 2020), Term Frequency-Inverse Document Frequency (TF-IDF) (Galán-García et al., 2014), Global Vectors for Word Representation (GloVe) (Akhter, 2019), and Word2Vec (Nandakumar, 2018) are a few of the frequently used word representation techniques that have been shown to increase the classification accuracy (Tarwani et al., 2019). Contextual expert knowledge is one of NLP's primary drawbacks. There are a lot of questionable claims regarding the ability to recognise sarcasm, for example, but how would one go about doing so when responding to a loss with a brief post like "Great game!" Thus, the issue is not language proficiency; rather, it is having knowledge pertinent to the discussion.

5. 2 Machine Learning in Cyberbullying Detection:

Another approach to cyberbullying detection that has been extensively employed by numerous researchers is machine learning based cyberbullying keywords. Furthermore, machine learning (ML), which is sometimes referred to as supervised, semi supervised, or unsupervised algorithms, is a subfield of artificial intelligence technology that allows systems to automatically learn and develop from experience without needing to be specially programmed (Dinakar, 2011). In supervised algorithms, a model that produces the intended prediction (i.e., based on annotated/labeled data) is constructed using multiple training instances. Unsupervised algorithms, on the other hand, are mostly used for clustering problems and are not dependent on data (Chen, 2012; Dinakar, 2011).

A model for identifying offensive comments on social networks and notifying the offending parties was put forth by Raisi and Huang (Hasan, 2023). This model has been trained using inflammatory comments from Ask.fm and X. Other writers (Buczak & Guven, 2015; Mccallum & Nigam, 2001) developed intelligent agent based communication systems that offer consoling emotional support to victims of cyberbullying. In order to identify aggressive patterns in user messages, Reynolds (Misra & Li, 2020) proposed a method for identifying cyberbullying in the social network "Formspring" that analyses offensive words and employs a threat rating system. In a similar vein, J48 decision trees achieved 81.7% accuracy.

The authors of (Harryzhang, 2011) outline the implementation of an online application that gives parents and school personnel in Japan the responsibility of identifying inappropriate content on unofficial secondary websites. Reporting instances of cyberbullying to federal authorities is the aim; 79.9% accuracy was attained in this work using SVMs. A Facebook framework has been proposed by Rybnicek (Joachims, 1998) to shield minors from sex teasing and cyberbullying. In order to track alterations in behaviour, the system analyses the content of images and videos as well as the actions of the user. (*(PDF) Brute Force Works Best Against Bullying*, 2017) used 3915 posted messages that were monitored from the Formspring.me website to create a list of offensive words. This study only achieved 58.5% accuracy [62](*(PDF) Brute Force Works Best Against Bullying*, 2017).

A methodology for distinguishing and categorising cyberbullying acts as harassment, flaming, terrorism, and racism is proposed by another study (Chatterjee, 2018). Because the author used a fuzzy classification rule, the accuracy of the results is lower (about 40%), but the classifier efficiency was increased by up to 90% when a set of rules was used.

A model for detecting cyberbullying based on sentiment analysis in Hindi-English code mixed language was created by authors in (Chatzakou, 2019). The Instagram and YouTube platforms served as the basis for the authors' experiments. The authors employ a hybrid model that outperforms eight baseline classifiers, with a f1-score of 82.96% and an accuracy of 80.26%.

Irena and Setiawan in their paper ‘ Fake News (Hoax) Identification on Social Media Twitter using Decision Tree C4.5 Method ’ (Irena & Setiawan, 2020) proposed applying supervised machine learning to a real world case of cyberbullying detection in X. Using two distinct feature extraction methods and a variety of machine learning algorithms, the study's Sequential Minimal Optimisation (SMO) classifier produced the best accuracy of all, 68.47 percent. An approach for detecting cyberbullying based on Instagram's social network has been proposed by the authors in (*(PDF) Detection of Cyberbullying Incidents on the Instagram Social Network*, 2020). The analysis of image contents and user comments served as the basis for the experiments. The findings demonstrate that the classification accuracy of linear SVM can be increased by employing multiple features; for example, by adding image categories as an additional feature, the accuracy of SVM increased from 0.72 to 0.78. In order to enhance online bullying, Nahar et al. (Ptaszynski, 2010) suggest developing a weighted directed graph model for cyberbullying that can be used to determine each user's predator and victim scores. They also suggest utilising a weighted TF-IDF scheme with textual features, such as second person pronouns and foul language.

A method for detecting hate content across several social media platforms is proposed by Salminen et al. (Dadvar, 2012). The authors utilised a total of 197,566 comments from four different platforms: Wikipedia, Reddit, YouTube, and X. Of these, 80% were classified as non hateful, and the remaining 20% were. Several machine learning algorithms were used in the experiments to test each feature independently and assess accuracy depending on feature selection. Dadvar et al. (Van Hee et al., 2015) proposed a suitable approach that combined roles typical of cyberbullying, content based, and user based, in addition to machine learning classifiers. Better performance was demonstrated by the results when all features were used simultaneously. A corpus of Dutch social media messages was developed by Van Hee et al. (*(PDF) Text Classification Algorithms: A Survey*, 2019) and annotated in various categories of cyberbullying, such as threats and insults. Additionally, the authors included a thorough account of the participants in bullying, including the identities of the victim, cyber predator, and bystander. By extending the insult, Zhao et al. (Sahlgren et al., 2018) were able to create bullying features through word embedding and achieve an f-measure of 0.78 using an SVM classifier. Furthermore, a dictionary of common terms used by neurotics in social networks was used to derive the novel features. The Word2Vec embedding model based neural network was employed by the authors in (*(PDF) Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media*, 2018) to represent textual health data with a semantic context.

Additionally, the Word2Vec model incorporates unique domain ontologies. Further information about a neural network model that can identify the semantic sense of uncommon words is given by these ontologies. The Bi-LSTM model is used to generate new semantic information that accurately separates structured from unstructured health data. To identify and categorise fake news on X, a different work is utilised: the decision tree C4.5 classifier based on the TF-IDF weighting method. To extract features for the recommended C4.5 classifiers, N-gram is also used (*Comparison between Naïve Bayes and Logistic Regression – DataEspresso*, 2017). A novel model that integrates the most pertinent documents, reviews, and tweets from news articles and social media has been proposed by authors in (Deb, 2016). They also developed a word embedding model that represented each word in a document with a low dimensional vector and a semantic meaning by integrating a topic2vec with Word2Vec. The authors also classified the data using machine learning (ML) and the previously mentioned models.

Since cyberbullying is regarded as a classification problem (i.e., classifying an instance as offensive or non offensive), this study has used a number of supervised learning algorithms along with physclogical safety climate theory (PSC) principles. Then they were further enhanced through classifiers and performance metrics to improve their performance and classification accuracy to yield the best performance score in identifying cyberbullying upon social media platforms, specifically on X. The following classifiers were used in the current investigation:

5. 3 Logistic Regression:

One of the well known methods that machine learning brought to the field of statistics is logistic regression (Snakenborg et al., 2011). Using the logistic function, the logistic regression algorithm creates a distinct hyper plane between two datasets (J. Patchin & Hinduja, 2010). Using features (inputs), the logistic regression algorithm forecasts based on the likelihood that a class will be appropriate for the input. Equation (1) states that, for example, the instance classification will be in the positive class if the likelihood is ≥0.5; if not, the prediction will be in the other class (negative class) (Tenenbaum, 2011). Logistic regression was employed in (Ang & Goh, 2010; Olweus, 2012) to implement cyberbullying prediction models.

$$\mathrm{h}_{\theta}\ (\mathrm{x})\ = \frac{1}{1+\ e^{-\theta^{T}x}},$$

(1)

If y = 1 (Positive class) and y = 0 (Negative class) respectively, then hθ (x) ≥ 0.5. According to (Barlińska, 2013), LR performs better with larger data sets and is effective for the binary classification problem. In order to reduce the error function, LR iteratively modifies the set of parameters (Barlińska, 2013).

5. 4 Logistic Light Gradient Boosting Machine:

Known as a gradient boosting framework that makes use of a tree based learning algorithm, LightGBM is one of the most potent boosting algorithms in machine learning (Ybarra et al., 2006). Still, it works better than CatBoost and XGBoost (Smith, 2008). LightGBM uses Gradient-based One-side Sampling (GOSS) to categorise the observations that are used to calculate the separation. The main benefit of the LightGBM is that it alters the training algorithm, which speeds up the process considerably (Raisi & Huang, 2016) and frequently results in a more effective model (Raisi & Huang, 2016; van der Zwaan, 2012).

Anomalies in big accounting data (Reynolds et al., 2011) and online behaviour detection (Muneer & Fati, 2020b) are two classification domains where LightGBM has been applied. However, the field of cyberbullying detection did not frequently employ LightGBM.

5. 5 Stochastic Gradient Descent:

An optimisation technique called stochastic gradient descent, or SGD, is used to determine the parameter values, or coefficients, of a function (f) that minimises the cost function (Rybnicek, 2013). Equation (2) states that SGD updates a parameter for every training example x(i) and label y(i).

$$\theta = \theta - \eta \cdot \nabla_{\theta} J\,(\theta; x^{(i)}; y^{(i)}), \tag{2}$$

As a result, SGD was applied in (Al-Garadi, 2016; Yin, 2009) to create cyberbullying prediction models on social networking sites. According to the authors in [82](Barlińska, 2013), SGD performs more quickly than NB and LR, but it does not have the minimum error that LR does.

5. 6 Random Forest:

Using average data to improve predictive accuracy and fitting control, the Random Forest (RF) classifier is an ensemble algorithm (Salminen et al., 2020) that matches multiple decision tab classifiers on various data sub samples (Dinakar et al., 2012). Ensemble algorithms combine multiple data classification algorithms, either of the same or different types(Ahlfors, 2010; Dadvar et al., 2013). The literature frequently used RF to create cyberbullying prediction models; studies by (Lenhart, 2010; Zhao, 2016) are among the examples. As a result, the variables for the classifier data are chosen at random using a number of trees that make up RF. The RF is constructed in the four streamlined steps that follow. N and M represent the number of examples (cases) and attributes in the classifier, respectively, in the training data.

- The number of examples (cases) in the training data is N, and the number of attributes in the classifier is M.

- A collection of arbitrary decision trees is generated by choosing random attributes. A training set is chosen for every tree by choosing n times from among all N instances that already exist. To estimate the error of the tree, the remaining instances in the training set are used to predict their classes.
- Each tree's nodes select M random variables upon which to base their decision. The training package uses certain m attributes to identify the most exceptional split. Unlike regular tree classifiers, which can be pruned after development, each tree is constructed from the ground up.
- A lot of trees are produced by this architecture. Those decision trees cast votes for the most common class. These procedures are called RFs. A model composed of multiple tree structured classifiers, with each tree voting for the most popular class, is constructed by RF (Salminen et al., 2020). The class with the most votes is the one that is chosen as the output.

5. 7 Adaboost:

Originally designed to increase the effectiveness of binary classifiers, adaptive boosting, also known as AdaBoost, is a popular ensemble learning technique (Havas et al., 2011; Webb et al., 2008). Iteratively learning from the mistakes made by weak classifiers, it builds stronger ones. For this reason, equal weights are initially assigned to each training observation. It makes use of multiple weak models and gives experimental misclassification observations higher weights. The accuracy of the misclassified observations increases as the results of the definitive boundaries acquired over multiple iterations are combined using multiple low models. As a result, the iteration's overall accuracy is improved (*Online Pestkoppenstoppen: Systematic and Theory-Based Development of a Web-Based Tailored Intervention for Adolescent Cyberbully Victims to Combat and Prevent Cyberbullying | BMC Public Health*, 2014). A similar dataset with two features and two classes is displayed in Figure 4 as an example of AdaBoost classifier implantation. Week learner #2 improves on a mistake made by weak leaner #1, and the accuracy of the misclassified observations is further improved when the two weak classifiers are combined (strong leaner).

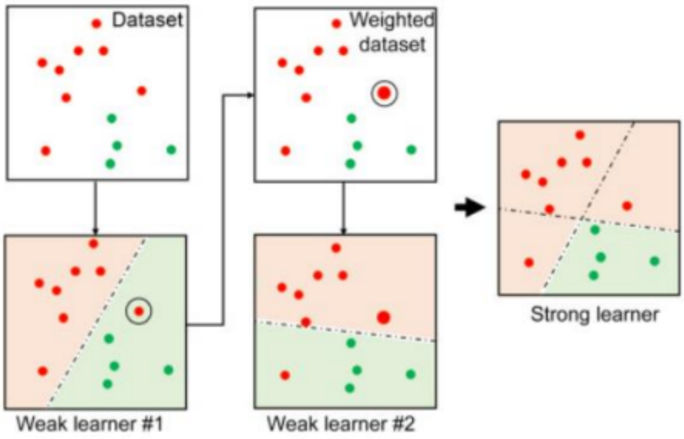


Figure 4 Implementation of Adaboost classifier (Havas, 2011)

Additionally, AdaBoost has been employed in the detection of cyberbullying by a few researchers, including (*KiVa Is an Anti-Bullying Programme | KiVa Antibullying Program | Just Another KiVa Koulu Site*, n.d.) and (Chatzakou et al., 2019), who used it to detect cyberbullying and achieved an accuracy of 76.39% using AdaBoost and features like unigrams, comments, profile, and media information.

5. 8 XGBoost

An open source, dependable gradient tree boosting model is called Extreme Gradient Boosting (XGBoost). Tianqi Chen began it as a research project in 2014 (T. Chen & Guestrin, 2016). It combines an ensemble of estimates from a set of trees as a supervised learning algorithm. When XGBoost is used instead of traditional gradient boosting decision trees, it can continue tree construction with missing values by converting them into a sparse matrix. This effectively helps prevent some overfitting issues. XGBoost also has the advantage of column sampling.
Given a dataset of form:

D = {(xi, yi) : i = 1...n, xi ∈ Rm, yi ∈ R} , It gets ‘n’ observations with m features each and with a corresponding variable y. Let ŷi be defined as a result given by an ensemble represented by the generalised model as follows:

$$\hat{Y}_i = \sum_{k=1}^{K} f_k\left(x_i\right) \tag{3}$$

In the above formula, fk is a regression tree, and fk (xi) represents the score given by the k-th to the i-th observation in data. Then the objective function to be minimized in step t is expressed as:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i,\ \hat{y}_i^{(t-1)} + f_t(x_i)) + \omega\left(f_t\right) \tag{4}$$

where ŷiˆ (t−1) denotes the prediction result of the previous t − 1 trees for sample xi , ft stands for the t tree, l is loss function and ω is the canonical term used for the t-th tree (Pan, 2018).

5. 9 Multinominal Naïve Bayes:

Multinomial Naive Bayes, or Multinomial NB, is a popular method for classifying texts and documents. However, NB was most frequently used to implement cyberbullying prediction models in the field of cyberbullying detection, as seen in (Hinduja & Patchin, 2008) and (Stauffer et al., 2012).

By using the Bayes theorem among features, NB classifiers were created. This model assumes that the text is generated by a parametric model that uses training data to find the model's Bayes-optimal parameter estimates. It classifies generated test data using those approximations (Notar et al., 2013). NB classifiers are capable of supporting an infinite number of distinct categorical or continuous functions. One dimensional kernel density estimation reduces a task of estimating high dimensional density, assuming that the functions are distinct. The NB algorithm is a learning algorithm that relies on the application of the Bayes theorem under strong (naive) independence assumptions. Consequently, NB was covered in great detail in (Fanti, 2012).

5. 10 Support Vector Machine Classifier:

A popular supervised machine learning classifier for text classification is called Support Vector Machine (SVM) (Joachims, 1998). After converting the original feature space into a higher dimensional, user defined kernel space, SVM looks for support vectors to maximise the margin (distance) between two categories. Initially, SVM approximates a hyperplane that divides the two groups. As a result, SVM chooses samples from both groups that are closest to the hyperplane; these samples are called support vectors (Ybarra & Mitchell, 2007).

The goal of SVM is to effectively discern between the two categories (positive and negative, for example). In the event that the dataset can be divided along nonlinear boundaries, the SVM implements particular kernels to properly rotate the function space. For a dataset that is difficult to separate, soft margin is used to reduce overfitting by assigning less weighting to classification errors along the decision boundaries (Havas, 2011). In this study, the basis function is an SVM with a linear kernel. The SVM classifier implementation for a dataset with two features and two categories is shown in Figure 5, where all training sample representations are represented as stars or circles. For each of the two categories in the training samples, support vectors—also called stars—are those that are closest to the hyperplane relative to the other training samples. Because two of the training results were on the incorrect side of the hyperplane, they were incorrectly classified.

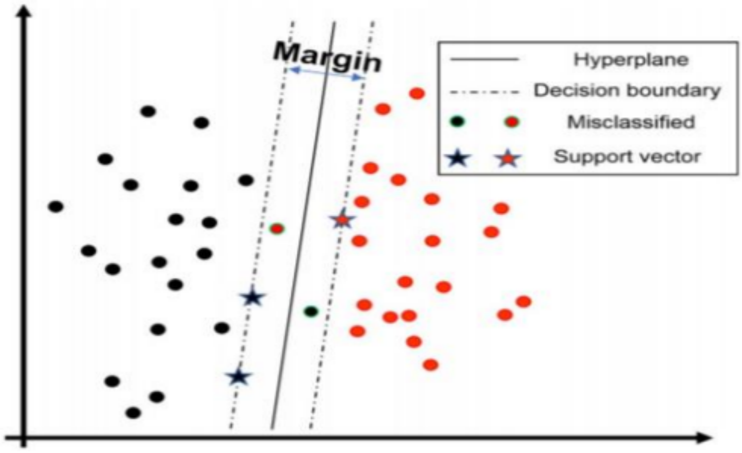


Figure 5 Implementation of Support Vector Machine (SVM) classifier (Jacobs, 2014)

Consequently, SVM was utilised in (*Lutte contre le harcèlement à l'école*, 2020) to build cyberbullying prediction models, and it was discovered to be successful and efficient. Nevertheless, the researach revealed that accuracy declined as data size increased, indicating that SVM might not be the best option for handling the frequent linguistic ambiguities that are characteristic of cyberbullying.

5. 11 Conclusion

The review of literature highlights the critical need for effective cyberbullying detection and prevention mechanisms due to the severe psychological impact on victims. Various international programs and strategies have been developed, emphasizing educational and psychological interventions, yet these are not universally implemented. Victims often seek anonymous support online rather than confiding in parents or teachers, pointing to the necessity of robust internet based solutions. Effective intervention approaches include increasing awareness, teaching emotional self management, and incorporating empathy training.

Technological approaches, particularly Natural Language Processing (NLP) and Machine Learning (ML), are at the forefront of cyberbullying detection. NLP methods focus on analyzing linguistic features across multiple levels, including phonology, morphology, lexical, syntactic, semantic, discourse, and pragmatic. Key techniques such as Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and sentiment analysis have demonstrated potential in identifying cyberbullying content.

Machine learning approaches, including supervised and unsupervised algorithms, offer promising results. Techniques like Logistic Regression, LightGBM, Stochastic Gradient Descent, Random Forest, AdaBoost, XGBoost, Multinomial Naïve Bayes, and Support Vector Machines have been utilized to classify and detect cyberbullying with varying degrees of accuracy. Each method has its strengths and limitations, often requiring a balance between computational efficiency and predictive performance. For this specific study Random Forest, Multinomial Naïve Bayes, XGBoost and Support Vector Machine Classifier has been chosen for the comparative analysis.

In summary, while significant progress has been made in developing detection and intervention strategies for cyberbullying, challenges remain. Effective solutions must integrate technological advancements with psychological insights to provide comprehensive support and protection for victims. Future research should continue to refine these methods, aiming for higher accuracy, contextual understanding, and broader application to ensure safer online environments.

6. Theory

In relation to the identification and remediation of cyberbullying, a number of theoretical frameworks have been investigated. According to Albert Bandura's **Social Learning Theory (SLT)**, people pick up behaviors from watching other people, especially from role models like parents, friends, and media figures (*Albert Bandura's Social Learning Theory In Psychology*, 2024). According to Bandura (1977), Behavior is learned through observation, imitation, and experience with the results of one's actions. SLT discusses how people might become aggressive online after seeing their peers act in this way and getting approval from their peers or social reinforcement when it comes to cyberbullying (Barlett, 2024). This theory emphasizes how social influence and behavior modeling play a major role in the proliferation of cyberbullying.

The **General Aggression Model (GAM)**, put forth by Anderson and Bushman (2002), is another pertinent framework. Using components from developmental psychology, social learning, and cognitive psychology, GAM explains how situational and individual factors affect aggressive behavior. It implies that through a cyclical process of reinforcement, being exposed to aggressive stimuli can cause an increase in aggressive thoughts, feelings, and behaviors. Because GAM takes into account both the internal state of the individual and external influences like media and social interactions, it is especially helpful in understanding the causes and effects of cyberbullying (Anderson & Bushman, 2002). This model offers a thorough framework for examining the complex psychological mechanisms underlying aggressive behavior and the external circumstances that intensify it, as well as the varied aspects of cyberbullying.

This study uses Dollard and Bakker's (2010) **Psychosocial Safety Climate (PSC) theory**, although SLT and GAM offered helpful insights. PSC is defined as the common understanding that the group values and guards its members' psychological comfort and safety. According to (M. F. Dollard & Bakker, 2010), it focuses on organizational policies, practices, and procedures that foster a supportive and safe environment, thereby lowering the prevalence of harmful behaviors like bullying. In contrast to SLT and GAM, which focus mainly on the situational and individual aspects of aggression, PSC highlights the role that the larger social environment plays in averting negative behaviors. This all encompassing strategy fits in nicely with the objectives of the study, which aims to create a safe and encouraging online environment in order to not only identify but also prevent cyberbullying. The incorporation of PSC principles into predictive models enhances their ability to consider psychosocial factors that impact cyberbullying, thereby facilitating the development of more efficacious detection and intervention tactics.

6. 1 Psychosocial safety climate theory

Psychosocial climate (PSC) refers to the specific element of an organization's overall climate that focuses on the policies, practices, and procedures that promote the psychological well being and safety of workers (Garrick et al., 2014). The primary factor influencing PSC is the quality of management and leadership within organizations. The PSC construct comprises four primary components (M. F. Dollard & Bakker, 2010) that are linked to established principles of best practice in the fields of stress prevention, intervention, and safety climate (Cheyne et al., 1998; Kompier & Kristensen, 2000). The first factor is the endorsement and dedication of senior management to promoting psychological safety, as demonstrated through their active involvement and commitment (Dollard, Tuckey, & Dormann, 2012). This aspect becomes apparent when senior management promptly and decisively take action to address and rectify issues that impact psychological comfort (Idris, Dollard, Coward, & Dormann, 2012). The second factor is the level of importance that the management places on the psychological health and safety of employees compared to productivity objectives (Hall et al., 2010). For instance, job demands, such as work pressure, can be adjusted to make them easier to handle. Management has the authority to provide various resources, such as work flexibility, autonomy, and social support, which can help alleviate demands and decrease work stress, ultimately benefiting the psychological protection and productivity of workers. The third aspect is organizational communication (Hall et al., 2010), which pertains to how the organization effectively communicates with employees regarding psychological health and safety matters that impact them, and ensures that employees are aware of these issues. The last component, organizational participation and involvement, refers to the active participation and consultation of all levels of the organization in stress prevention. This includes the integration of stakeholders such as employees, unions, and health and safety representatives in the processes related to occupational (psychological) health and safety (Idris et al., 2012).

Multiple studies employing multilevel models have provided empirical evidence supporting the hypothesis that PSC (positive school climate) decreases instances of bullying. In a study conducted by Bond et al. (2010), it was discovered that the presence of Police Social Capital (PSC) at the police station level was a reliable predictor of workplace bullying within the following 12 months. Law, Dollard, Tuckey, and Dormann (2011) conducted cross sectional multi level research and discovered a negative correlation between organizational level perceived social climate (PSC) and workplace bullying. Furthermore, they found that workplace bullying was associated with psychological distress. The significance of these findings lies in their ability to demonstrate that knowledge of shared perceptions of PSC can serve as a predictor for bullying behavior (Bond et al., 2010; Law et al., 2011). The PSC theory has traditionally focused on shared perceptions of PSC. However, recent theorization suggests that it is worthwhile to investigate psychological PSC, which refers to individual perceptions of PSC, rather than shared perceptions. This is particularly important when examining the subsequent perceptions of effort reward imbalance and their impact on psychological health (Owen Bailey, & Dollard, 2016). Research conducted at the individual level has demonstrated that the escalation of bullying and its outcomes are influenced by the victim's perception of the context in which the bullying occurs, particularly in terms of their ability to express themselves and take action to address the bullying (Kwan Tuckey, & Dollard, 2016). This study examines the impact of individual perceptions of PSC (Perceived Social Control) and organizational procedures on cyberbullying.

6. 2 How PSC relates to cyberbullying

The hierarchy of controls is a fundamental principle in occupational health and safety intervention. It suggests that identifying the underlying cause, by recognizing more distant causes, will result in a more successful and economical control or intervention strategy (M. Dollard & Mcternan, 2011). Within the realm of social media, the concept of addressing cyberbullying can be approached using a comparable principle. By identifying and addressing the underlying causes of cyberbullying, rather than solely focusing on the observable signs, it is possible to implement more impactful interventions.

Psychosocial safety climate (PSC) is an underlying factor that contributes to social media bullying, serving as a root cause or a factor that influences other causes. Continuous exposure to negative psychological social conditioning (PSC) on social media platforms can ultimately lead to adverse health consequences for users. The transmission process by which PSC influences cyberbullying is determined by the way PSC is implemented. Espoused PSC refers to the stated actions that platform administrators claim they will take to address cyberbullying. On the other hand, enacted PSC refers to the specific measures that have been put into practice. Although psychological health and safety policies are influenced by PSC, they are more distant in terms of their impact. On the other hand, the procedures or mechanisms for implementing these policies into daily social media interactions (enacted PSC) are closer to the site of change and are therefore the main focus of this research.

Until now, there has been limited discussion in the literature about the specific actions that result from climates, except for some instances in the safety climate literature (Zohar & Luria, 2005). This study also adds to the existing body of research on the climates of organizational and social media platforms.

The correlation between the implementation of PSC (Public Service Commission) and cyberbullying can be comprehended through three psychosocial mechanisms: (1) a climate of mistreatment, (2) the design of work, and (3) the escalation of conflict. The first mechanism, known as mistreatment climate, refers to a specific social media platform climate that is characterized by mistreatment. According to scholars, the safety climate can influence how people address the problem of bullying and the extent to which platform administrators are willing to take action in the future (Salin, 2008). Expanding upon this concept, PSC is suggested as a distinct type of safety climate that is productive in its nature, resulting in the development of a particular climate of bullying and mistreatment. Put simply, PSC refers to a wide ranging concept involving mistreatment that can lead to the development of more specific mistreatment climates (Einarsen, Skogstad, Rørvik, Lande, & Nielsen 2018).

Based on Vroom's expectancy theory of 1964 (Yang et al., 2014) , Bandura's social learning theory of 1986 (Oliverio, 2023), and Katz and Kahn's role theory (Flood, 2017), the concept of mistreatment climate explores the extent to which bullying is considered unacceptable in a given environment. This is closely related to the concept of safety climate, as discussed by Zohar and Luria (2005). The current climate of mistreatment provides users with information regarding their expectations of how their behavior will lead to certain outcomes, as well as the desired behavior they should exhibit in response to mistreatment. Users gain an understanding of the likely outcomes that may arise from the alignment or mismatch between the climate and their behavior in their respective roles. Regarding the issue of mistreatment in the climate, judgments are made regarding the level of tolerance, reward, or punishment for bullying.

(Yang et al., 2014) conducted a meta analysis to examine the impact of various mistreatment climates on users' motivation and behaviors related to mistreatment. The researchers made predictions and discovered evidence supporting the connection between the mistreatment climate and role behavior. They found that the climate had an impact on motivation, such as prevention motivation, as well as behavior, such as compliance with measures and participation, in relation to mistreatment. Similarly, the safety climate literature provides supporting evidence from meta analytic reviews that establish a relationship between safety climate, safety motivation, and behavior (Christian, Bradley, Wallace, & Burke, 2009). Regarding cyberbullying, (Baillien, 2013) discovered that anti-bullying policies have a significant negative correlation with bullying, independent of organizational change (which has a positive correlation) and people oriented culture (which has a negative correlation). Similarly, the implementation of anti-bullying policies has been found to be associated with reduced incidences of bullying (Cooper-Thomas et al., 2013). When considering these arguments collectively, it is anticipated that PSC will create an environment specifically focused on mistreatment through bullying, which will be evident through the implementation of anti-bullying measures that impact the occurrence of bullying. Therefore, these anti-bullying procedures can be considered as a manifestation of implemented psychological safety culture (PSC).

6. 3 Aligning PSC Mechanisms with Predictive Analytics Models

Work design is another psychosocial mechanism that explains the connection between PSC (perceived social competence) and bullying. The PSC, which is greatly influenced by senior managers, serves as an indicator of how much senior management values the psychological safety of workers. PSC is a guiding principle that determines how work is structured and the level of work excellence that is attainable. In short, PSC influences job design. Within the realm of predictive analytics models for cyberbullying detection, this means that the principles that form the basis of PSC can provide guidance for the design and characteristics of these models. For instance, one may develop features that evaluate the caliber of online interactions and the psychological consequences of these interactions based on PSC principles. The work design hypothesis proposed by Leymann in 1996 suggests that bullying is influenced by job design factors (Teixeira, 2016). The translation of this concept into the digital domain involves analyzing how the design elements of social media platforms, such as the user interface and interaction rules, impact the prevalence of cyberbullying. A recent systematic review found that role conflict, workload, role ambiguity, job insecurity, and cognitive demands are important factors that contribute to bullying in the workplace (Van den Brande, Baillien, De Witte, Van der Elst, & Godderis, 2016). Some studies have also reported that these factors can have the opposite effect, meaning they can reduce the likelihood of bullying (Hauge, Skogstad, & Einarsen, 2011). These factors can be applied to social media environments, where the clarity of user roles, the amount of interaction, and the perceived level of security can be used to create models that predict the risks of cyberbullying.

In addition to this, when senior management values productivity more than the welfare of workers, it can lead to the encouragement and promotion of bullying by middle managers and first line supervisors. This can be achieved by imposing higher work pressure and heavier workloads on their subordinates in order to complete tasks (Bailey, Dollard, & Tuckey, 2014; Ceja, Escartín, & Rodríguez-Carballeira, 2012). In the same vein, social media platforms that prioritize engagement metrics at the expense of user security may unintentionally create environments that are more conducive to cyberbullying. Predictive models can include metrics that represent this prioritization equilibrium, such as the speed at which reports of cyberbullying are addressed and the implementation of proactive monitoring.

Challenging work conditions can also deplete both job and personal resources, which would otherwise be beneficial in handling job demands (according to Conservation of Resources Theory, Hobfoll, 2001) and enduring bullying (Tuckey & Neall, 2014). Research indicates that job stressors, specifically workload, are a significant predictor of decreased psychological detachment (Sonnentag & Fritz, 2015). Furthermore, reduced detachment has been found to result in heightened strain when dealing with bullying (Moreno-Jiménez, 2009). Within the realm of cyberbullying, extensive online participation and exposure to adverse interactions can have a comparable effect of diminishing users' psychological resilience. Predictive models can utilize engagement metrics and sentiment analysis to accurately detect individuals who are vulnerable to either being victims or perpetrators of cyberbullying.

In addition, demanding occupations hinder the attainment of objectives and are likely to result in feelings of frustration, as noted by Karasek (1979) and supported by various articles such as (Crawford & Detar, 2023). Furthermore, this adverse impact can potentially contribute to the development of aggressive and bullying behaviors in the workplace. (Harris, Harvey, Harris, & Cast, 2013) demonstrated a positive correlation between job frustration and employees' inclination to mistreat their colleagues. When it comes to social media, feeling frustrated with how the platform works or not having your social expectations met can cause people to behave aggressively online. Predictive models can incorporate attributes that identify indications of annoyance and hostility in user interactions to proactively identify potential instances of cyberbullying.

Inadequate work design provides a conducive environment for bullying to occur (Salin, 2003). Therefore, by implementing work redesign procedures, PSC can have an impact on the exposure to bullying. For social media platforms, it is imperative to integrate PSC principles into platform design and management practices in order to establish a more secure online environment. To enhance the accuracy of cyberbullying detection and mitigation, predictive analytics models should incorporate these principles. This will help ensure that interventions are in line with the objective of maintaining a psychologically healthy and safe user experience.

Additionally, PSC may be associated with bullying by means of a third psychosocial mechanism known as the conflict escalation hypothesis (Zapf & Gross, 2001). This hypothesis suggests that if social conflicts on the platform, characterized by negative interpersonal relationships, are not addressed, they have the potential to escalate into bullying. Unclear job responsibilities and conflicting objectives can lead to a competitive environment with low trust. Additionally, high work pressure often suggests that organizations do not prioritize conflict resolution, resulting in limited time and attention given to resolving conflicts (Zapf & Einarsen, 2005). Psychosocial health and safety is a wide ranging concept that encompasses various aspects of prosperity. In this context, PSC (Psychosocial Safety Climate) should play a role in determining how conflicts are handled on social media platforms. In their study, (Einarsen et al., 2016) introduced the notion of a climate for conflict management, which pertains to employees' evaluations of an organization's conflict resolution procedures. This concept is considered a subset of PSC. Essentially, they suggested that PSC generates protocols that result in equitable and foreseeable exchanges between supervisors and staff members. The researchers discovered a notable adverse correlation between the climate for conflict management and bullying. They hypothesized that this correlation was attributable to the limited range of the specific aspect of the climate. It is anticipated that organizations with high PSC (Psychological Safety Climate) would have established conflict resolution procedures (a type of enacted PSC) to promptly address conflicts before they escalate into bullying (Escartín Solanelles et al., 2013).

Overall, high PSC social media platforms decrease cyberbullying by utilizing three psychosocial processes: (a) creating an environment that discourages bullying mistreatment, (b) designing the platform in a way that minimizes conflict, and (c) implementing effective conflict resolution strategies. This is demonstrated by the implementation of protocols to (i) tackle cyberbullying, (ii) alleviate stress through user experience design, and (iii) aid in digital conflict resolution, which is known to be challenging.

Regarding the connections between cyberbullying, psychological health outcomes, and psychosocial safety climate (PSC), this study makes a number of assumptions. First of all, PSC has a positive correlation with PSC that has been put into practice, such as policies to deal with bullying on social media and initiatives to use UX design to lessen stress. Second, it implies that there is a negative correlation between the prevalence of cyberbullying and implemented PSC. Finally, it suggests that the negative relationship between PSC and cyberbullying is mediated by enacted PSC and that the beneficial effects of cyberbullying on psychological health issues like depression, psychological distress, and emotional exhaustion have been lagging behind. Putting all of this together, a common belief is that PSC is inversely and negatively correlated with poor psychological health outcomes via the mediation of bullying exposure and enacted PSC.

In this study there is a presumption that perceptions of PSC are related to the objective manifestation of PSC. However, individuals may have different experiences of PSC due to variations in leader member exchanges, as described in the leader member exchange theory of leadership (Graen & Uhl-Bien, 1995). Furthermore, people's understanding of climate is formed by assessing particular aspects of the environment based on their importance to personal values and contentment (Griffin & Neal, 2000; Neal & Griffin, 2006). Understanding of procedures may indicate objective variations in implemented PSC, as well as subjective variations among individuals. The psychosocial mechanisms described above explain how the enacted PSC procedures are connected to a decrease in bullying. These mechanisms theoretically result in a significant reduction of risk factors associated with bullying. Having a comprehensive understanding of procedures can also lead to awareness of the psychological hazards associated with bullying and its underlying causes. It is anticipated that this knowledge of the procedures as a whole will have an impact on behavior, resulting in a decrease in risks, including instances of bullying.

Furthermore this study will try to enact some PSC principles into cyberbullying detection model and see whether it creates any difference in terms of accuracy and processing time.

6. 4 Operationalization

This study will adopt the psychological safety climate (PSC) theory principles to explore the effectiveness of various predictive analytics models in detecting cyberbullying. By leveraging sentiment analysis, it aims to examine the social relevance and impact of these models, specifically focusing on random forest, GXBoost, support vector classifier, and Multinomial Naive Bayes.

The study will operationalise the comparison of these predictive models through several key aspects:

1. **Model Selection and Training:** Each predictive model will be selected based on its unique algorithmic approach to classification and prediction. These models will be trained on a dataset containing postive and negative words. The training process will involve tuning hyperparameters to optimize performance.

2. **Sentiment Analysis Integration:** To understand the context and sentiment behind the content, sentiment analysis will be performed on the data. This will help in assessing the emotional tone and potential impact of the content, which is crucial for detecting cyberbullying. Sentiment scores will be integrated into the predictive models as features to enhance their accuracy.

3. **Performance Metrics:** The study will evaluate the predictive models using several performance metrics, including accuracy, precision, recall and F1 score. These metrics will provide a comprehensive understanding of each model's ability to detect cyberbullying accurately.

4. **Comparative Analysis:** A comparative analysis will be conducted to highlight the strengths and weaknesses of each model. This analysis will consider factors such as computational efficiency, scalability, and ease of implementation in real world scenarios.

5. **Social Impact Assessment:** To align the study with social relevance and with PSC principles, the implications of using these predictive models will be assessed for the cyberbullying detection in different social contexts. This includes evaluating how well the models can adapt to various linguistic and cultural nuances present in the data.

6. **Ethical Considerations:** Ethical considerations will be taken into account, particularly concerning data privacy and the potential for bias in predictive analytics. The study will ensure that the models are developed and tested with a focus on fairness and ethical use.

7. **Stakeholder Involvement:** Input from stakeholders, such as educators, parents, and online platform moderators, will be gathered to understand the practical requirements and challenges in deploying these models for cyberbullying detection. Their feedback will be used to refine the models and make them more applicable in real world settings.

By operationalising the study in this manner, the research aims to provide a thorough comparative analysis of predictive analytics models in the context of cyberbullying detection. This approach will not only highlight the technical capabilities of the models but also their social relevance and potential impact on mitigating cyberbullying.

7. Methodology:

After discussing in theory chapter about how the theory (PSC) relates to cyberbullying and how it can be alligned with the predictive machine learning models. This chapter will detail the methodological approach used in this research to provide a comprehensive overview of the proposed model and the data description. Furthermore, it will explain the methods used for the comparative analysis.

7. 1 Model Overview

Figure 6 illustrates the proposed model of cyberbullying detection, where mainly it has six phases: the preprocessing phase, the feature extraction phase, random over sampling phase, classification phase, hyperparameter tuning and then the evaluation phase. Each phase has been discussed in detail in this section.

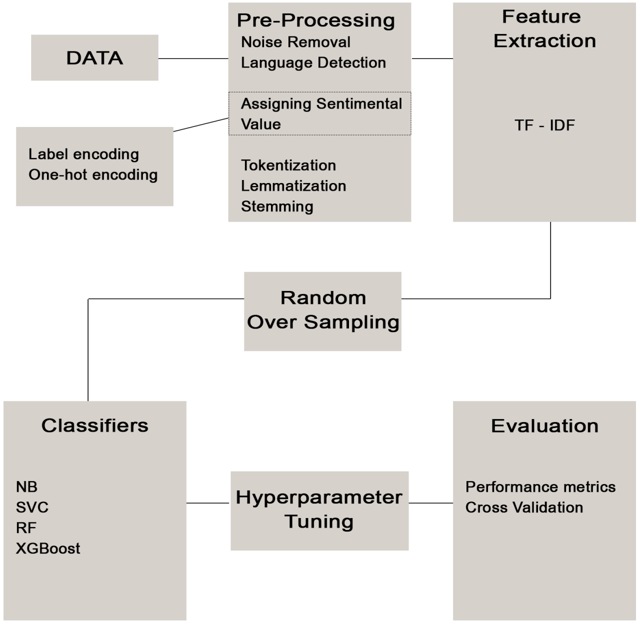


Figure 6 Proposed Model Overview

7.1.1 Overview and structure of model methodology:

The methodology of this experimentation is based on four steps, in the first step the data is collected from **UCI, Computer Science department** consisting of scrapped tweets by the users on X. In the second, the collected data is preprocessed before the classification from the noise. In the third step, the data is classified by using machine learning (ML) algorithms, deep learning, 5-fold cross validation results and by Hyperparameter Tuning – GridSearchCV – Random Forest. In the final step, the performance of the models is evaluated.

7. 2 Dataset description:

Data on social networking sites is considered confidential, and users of social media occasionally exhibit reluctance in disclosing personal data. Most studies gather data from common online communities such as X, Facebook, and YouTube. The datasets are created separately by using a freely accessible API or by extracting data from websites. Formspring is a highly utilized dataset that has undergone updates over time. According to (Reynolds et al., 2011), Formspring had approximately 4000 samples in its initial year, but that figure has now tripled. Among various authors, the only datasets that have been reused are Kongregate, Slashdot, and MySpace, which are available in multiple versions. The original datasets exhibit a notable disparity, as the majority of studies utilize datasets where less than 20% of the samples are categorized as cyberbullying. This imbalance poses a challenge as it has been extensively demonstrated to affect the predictive abilities of machine learning classifiers (Chawla, 2005). Multiple studies (Al-Garadi et al., 2016; Huang et al. 2014b; Roy et al., 2022) have employed synthetic oversampling or under sampling methods to create a more evenly distributed dataset, leading to improved classification outcomes. This is due to the inherent asymmetry of cyberbullying. Research on cyberbullying is expected to yield non normal distributions, as indicated by (Bauman et al., 2012).

It is worth mentioning that most of the datasets are classified as cyberbullying, even though there is no single post or message present. Given the repetitive nature of cyberbullying, it is improbable to identify a cyberbullying incident based on a single text message. Recognizing a pattern of consistently aggressive online messages targeting a particular individual is crucial for accurately defining cyberbullying. In summary, it is contended that the current datasets form the foundation of the entire study in this field. Therefore, it is imperative to initiate a paradigm shift in the way data portrays cyberbullying to facilitate more comprehensive research in the future. Furthermore, it is crucial to comprehend that the language is highly intricate, encompassing the utilization of specific words that possess diverse connotations in varying contexts and when conveyed with different intentions, expressions, tones, and motives. Therefore, (Saini et al., 2023) offers a concise analysis in which they explore the reasons why certain words are present in both cyberbullying and non-cyberbullying scenarios:

"

1. Ambiguity of Words: A few words can be used in several different ways. They could be utilized kindly in one situation and nefariously in another. As an illustration, while "stupid" might be casually used among friends without any malice intended, it can also be used to degrade and ridicule someone when engaging in cyberbullying.

2. Common Vernacular: Certain terms are part of everyday language and are used in a variety of circumstances. These words can appear in both good and bad contexts without denoting any particular bullying action. Words like "friend," "post," and "comment" are commonly used in both cyberbullying and non cyberbullying contexts.

3. Emotional Expressions: Emotions are an important element of human communication, and certain strong emotions can be expressed in different settings using similar terms. Words such as "angry," "sad," and "happy" may appear in both cyberbullying and non cyberbullying content since they are used to describe emotions in a variety of circumstances.

4. Contextual Factors: The context in which a word appears strongly influences its meaning. For instance, the word "help" can be employed in both cyberbullying and non cyberbullying contexts. In a non cyberbullying context, it could refer to someone seeking or offering assistance, however in a cyberbullying context, it could be used sarcastically to tease or criticize someone.

5. Neutral Terms: Some words are completely neutral, with no positive or negative meanings. Because they lack specific destructive or supporting meanings, these words can exist in both cyberbullying and non cyberbullying texts.

6. Collateral Usage: In some cases, both bullies and bystanders seeing or reacting to cyberbullying situations may use the same terms. Similar rhetoric may be used by spectators to criticize bullying or express empathy for the victim.

”

The dataset used for this research comprises approximately 2000 scraped tweets, which are labeled based on their location, text content, and the date and time of their creation/posting. The figure 7 below depicts the word cloud prior to (on the left) and subsequent to (on the right) text preprocessing.

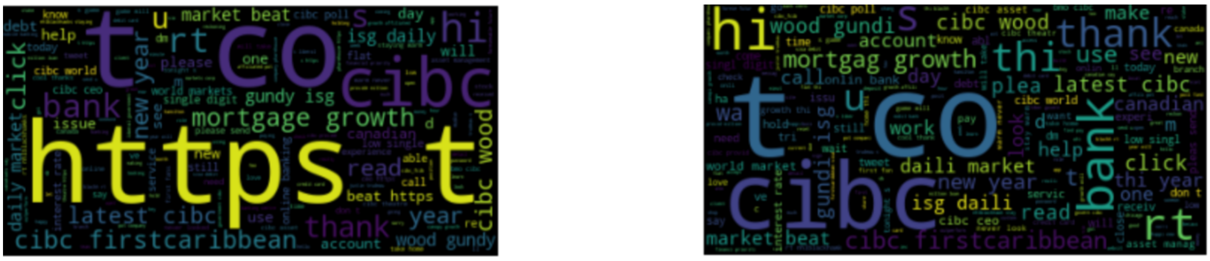


Figure 7 World cloud

In the original tweets ‘https’ was a common occurrence but is no longer a common occurrence after text pre processing while some of the common word occurrences in tweets after text pre processing are ‘thank’, ‘co’, ‘bank’ and ‘cibc’.

7. 3 Sentiment Analysis

Sentiment analysis, sometimes referred to as opinion mining, is a natural language processing (NLP) method for figuring out the text's emotional undertone. It entails examining textual data to locate and extract subjective data, including the writer or

speaker's opinions, attitudes, feelings, and sentiments.
Important facets of sentiment analysis include:

1) Polarity
   - • Positive: Expresses a feeling of favorability.
   - • Negative: Expresses an adverse opinion.
   - • Neutral: Expresses an impartial or neutral attitude.
2) Intensity: Sentimentality can be described as mildly positive, very positive, etc.
3) Emotion Detection: Sentiment analysis is capable of classifying text into distinct emotions such as happiness, surprise, anger, sadness, and more, in addition to positive and negative ones.
4) Aspect based Sentiment Analysis: This technique entails assessing feelings regarding particular facets or characteristics of a good, service, or subject that is discussed in the text. For instance, recognizing feelings about various elements such as cost, effectiveness, and quality in a product review.

Only the first component of sentiment analysis—polarity—has been evaluated in this study. This implies that a tweet can be classified as neutral, positive, or negative. This was only done to connect the research to sentimental score, one of the PSC tenets. The sole purpose of the analysis was to determine the sentiment polarity on social media sites like X, allowing the administrators to better understand their PSC and take focused measures to improve the psychological health and safety of their online environment from cyberbullying.

The following steps shows how the implementation of it was done in the dataset:

1. **Data Collection**:

   - Gathered text data from various sources on the platform, such as user posts, comments, reviews in terms of tweets

2. **Sentiment Analysis**:

    - The dataset contains approximately 2000 different (scrapped) tweets with the following attributes:

        - ***'id'*** : unique 19 digit id for each tweet
        - '***created_at'*** : date & time of each tweet (or retweet)
        - ***'text'*** : tweet details/ description
        - ***'location'*** : origin of tweet

    - Applied sentiment analysis tools to categorize the text data into positive, negative, and neutral sentiments.

3. **Assessing PSC**:

    - **Positive Sentiment Analysis**:

        - Look for comments that reflect a positive user experience, support from the platform's moderation team, recognition of contributions, and overall satisfaction. These indicate a strong PSC.

    - **Negative Sentiment Analysis**:

        - Identify comments that mention harassment, cyberbullying, lack of support, and unresolved conflicts. These indicate areas where PSC needs improvement.

    - **Neutral Sentiment Analysis**:

        - Analyze comments that provide neutral feedback or indicate indifference. These can point to policies or practices that may need more attention to increase their effectiveness and engagement.

4. **Actionable Insights**:

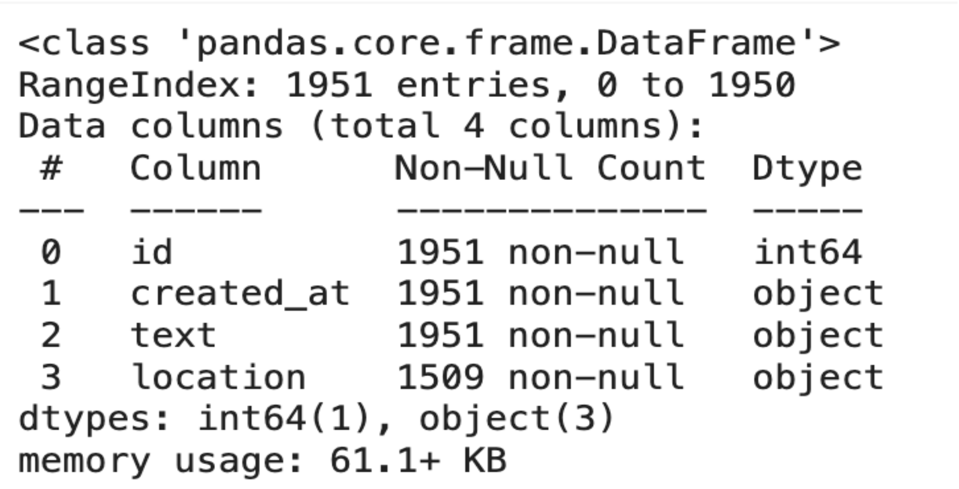
    - Use the sentiment analysis results to inform PSC-related strategies for the platform. For example, if negative polarity is high concerning bullying, the platform can implement or strengthen anti bullying measures and support systems.
    - Monitor changes over time to assess the impact of interventions aimed at improving PSC.

7. 4 Data preprocessing

The tweets should be cleaned and preprocessed before classification. Data preprocessing is carried out in four steps using Natural language processing (NLP) and tweet processing also shown in Figure 21: (Data cleaning, checking duplicates, stopword removal), tokenization, lemmatization and stemming. Each step is further discussed below:

7. 5 Data Reading

The csv file (tweets.csv) containing the data was called to shape the data in the form of rows and columns. Then the data information was fetched as shown below in figure 8.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1951 entries, 0 to 1950
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   id          1951 non-null   int64 
 1   created_at  1951 non-null   object
 2   text        1951 non-null   object
 3   location    1509 non-null   object
dtypes: int64(1), object(3)
memory usage: 61.1+ KB
```

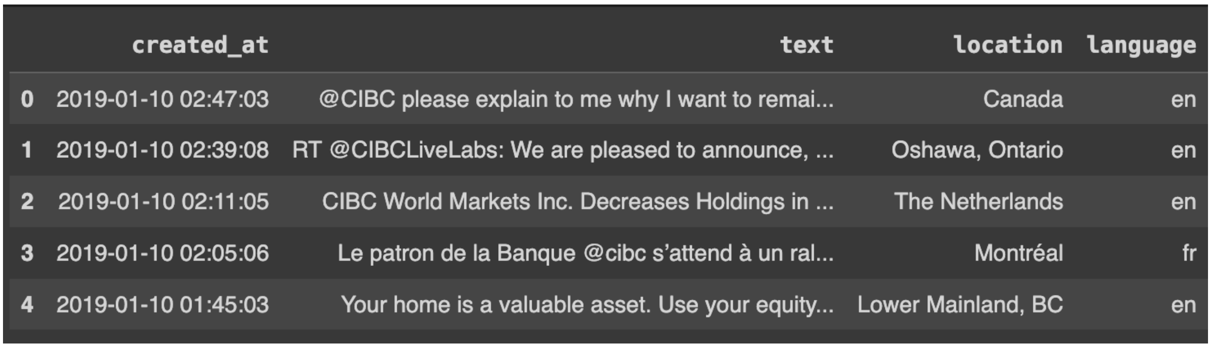
Figure 8 Data Reading Information

The information states that ‘id' is a 16 digit random number associated with each tweet. This should have no relevance on final modeling & can be dropped. 'created_at' is of object type & will be converted to datetime format. It also shows that the 'location' has 442 missing records.

7. 6 Data cleaning

There are two kinds of noise: external noise (such as URLs, hashtags, tweet sizes, etc.) and internal noise (content). The first step in cleaning up a dataset is to remove any items that introduce noise. We call this procedure "data cleaning." Using the preprocessor library of tweet preproccessor, special characters (&, +, /, %), numbers, single characters and user mentions (@user), tags, links, whitespaces, duplicate tweets, retweets, and tweet images are all eliminated. Clean tweets are converted to lowercase after being fed through the preprocessor with all of the tweets. The following is an example of a tweet before and after data cleaning:

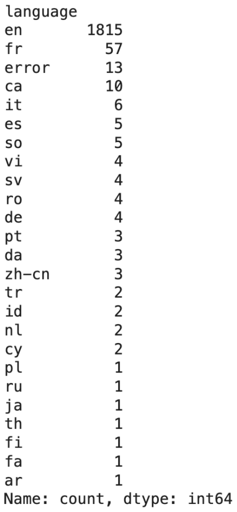
7.6.1 Language detection



	created_at	text	location	language
0	2019-01-10 02:47:03	@CIBC please explain to me why I want to remai...	Canada	en
1	2019-01-10 02:39:08	RT @CIBCLiveLabs: We are pleased to announce, ...	Oshawa, Ontario	en
2	2019-01-10 02:11:05	CIBC World Markets Inc. Decreases Holdings in ...	The Netherlands	en
3	2019-01-10 02:05:06	Le patron de la Banque @cibc s'attend à un ral...	Montréal	fr
4	2019-01-10 01:45:03	Your home is a valuable asset. Use your equity...	Lower Mainland, BC	en

Figure 9 Language Detection of Tweets

After checking upon tweets language as tabularly described above in figure 9, the flitered dataset only contained english texts that were around 1815 tweets as shown in the figure 10 below

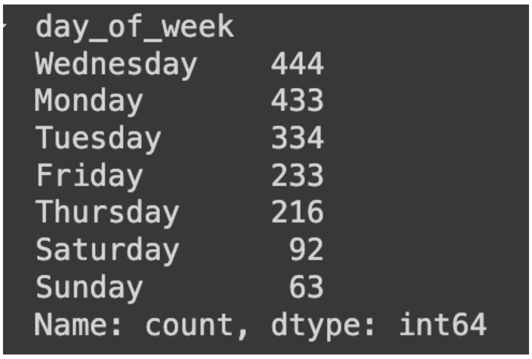


```
language
en       1815
fr         57
error      13
ca         10
it          6
es          5
so          5
vi          4
sv          4
ro          4
de          4
pt          3
da          3
zh-cn       3
tr          2
id          2
nl          2
cy          2
pl          1
ru          1
ja          1
th          1
fi          1
fa          1
ar          1
Name: count, dtype: int64
```

Figure 10 Filtered Dataset

7. 7 Assigning sentiment value

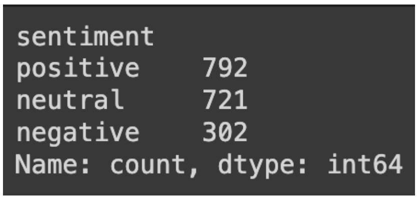
After data cleaning and language detection, the tweets were aligned according to the time and day of the week they were posted as shown in the figure 11 below, most of the tweets were posted on Wednesday (middle of the week) followed by Monday while least number of tweets were posted on weekends.



```
day_of_week
Wednesday    444
Monday       433
Tuesday      334
Friday       233
Thursday     216
Saturday      92
Sunday        63
Name: count, dtype: int64
```

Figure 11 Number of Tweets on Day_of_week

Then a function was called to analyse and assign sentiment value to the tweets in terms of positive, negative or neutral. Upon giving the count command, majority of the tweets were positive followed by neutral and the least were tweets having negative sentimental value. The exact values are shown in the figure 12 below:



```
sentiment
positive    792
neutral     721
negative    302
Name: count, dtype: int64
```

Figure 12 Number of Count of assigned sentiment Tweets

Plotting all this into a bar chart in figure 13, it was analysed that all weekdays (except Thursday) have majority of tweets as positive tweets, while weekends and Tuesdays have majority of tweets as neutral tweets. Wednesday has a highest number negative tweets.

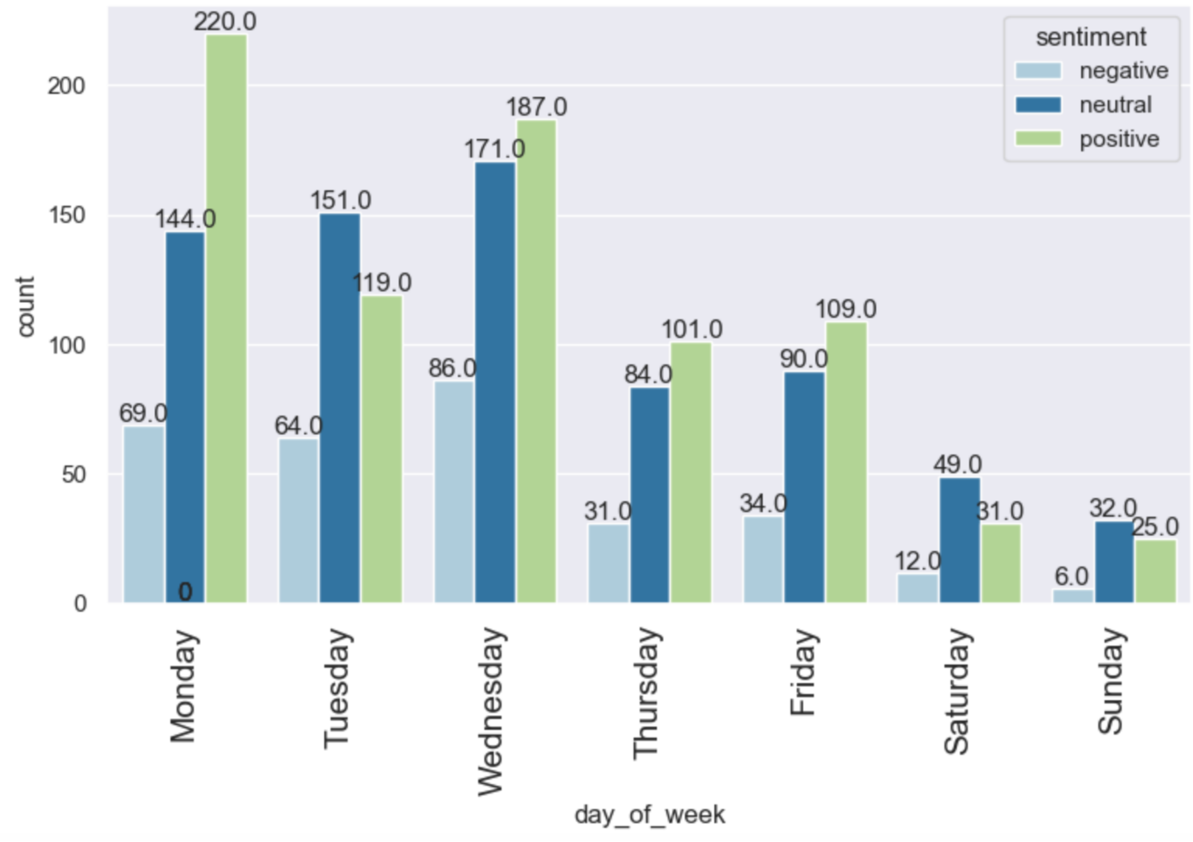


Figure 13 Bar Chart representation of Daily Sentimental Tweets

Furthermore, figure 14 presents the number of tweets posted as per time of the day, majority of the tweets were posted in late afternoons and evenings peaking after 3pm.

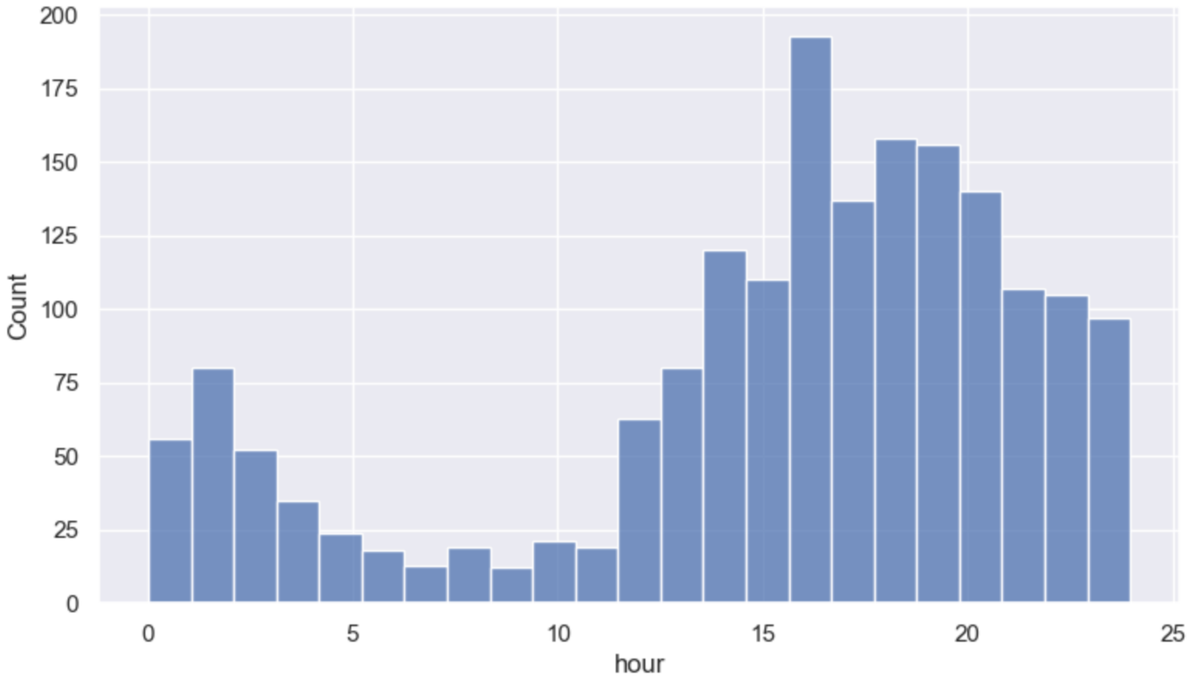


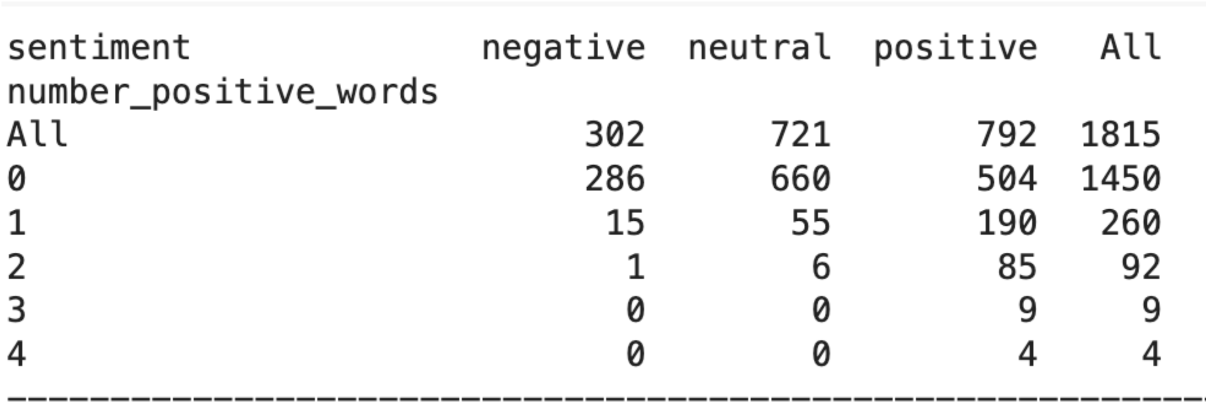
Figure 14 Number of tweets posted as per time of the day

Prior to tokenization, it is crucial to tally the occurrences of positive and negative words and subsequently compare them with the approved list of positive and negative words, as elaborated in the following section.

7. 8 Positive_words & negative_words

To create new feature and to access data more precisely for possible cyberbullying, number of positive and negative words in each tweet were counted. The list of all positive and negative words were borrowed from **University of California, Computer science department database**. This list was compiled over many years starting from their first paper published in 2004 (Hu and Liu, KDD-2004).

The following were the results when the number of positive words and negative words per tweet were counted in respect to their sentiments



```
sentiment               negative  neutral  positive   All
number_positive_words
All                          302      721       792  1815
0                            286      660       504  1450
1                             15       55       190   260
2                              1        6        85    92
3                              0        0         9     9
4                              0        0         4     4
----------------------------------------------------------
```

Figure 15 Numerical presentation of Sentiments as per number of count for Positive Words

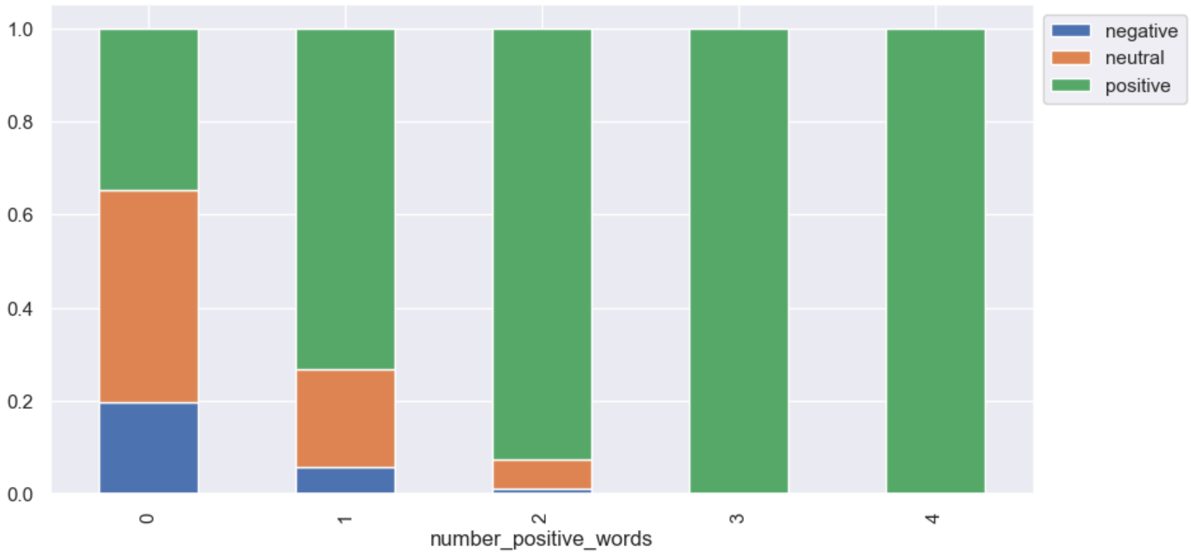
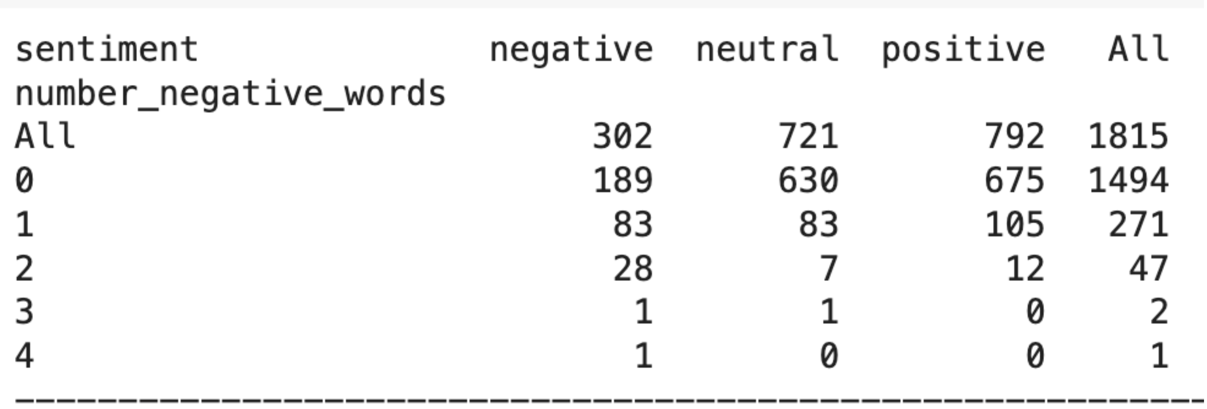


Figure 16 Bar chart presentation of Sentiments as per number of count for Positive Words

Figure 15 and Figure 16 depict the distribution of positive words in tweets. Figure 15 is presented as a numerical chart, while Figure 16 is displayed as a bar chart. The charts show that tweets with 0 and 1 count for the number of positive words have all three sentiments: 'neutral', 'positive', and 'negative'. However, tweets with 1, 2, 3, and 4 positive words are predominantly associated with a 'positive' sentiment, as anticipated.



```
sentiment              negative  neutral  positive   All
number_negative_words
All                         302      721       792  1815
0                           189      630       675  1494
1                            83       83       105   271
2                            28        7        12    47
3                             1        1         0     2
4                             1        0         0     1
------------------------------------------------------------
```

Figure 17 Numerical presentation of Sentiments as per number of count for Negative Words

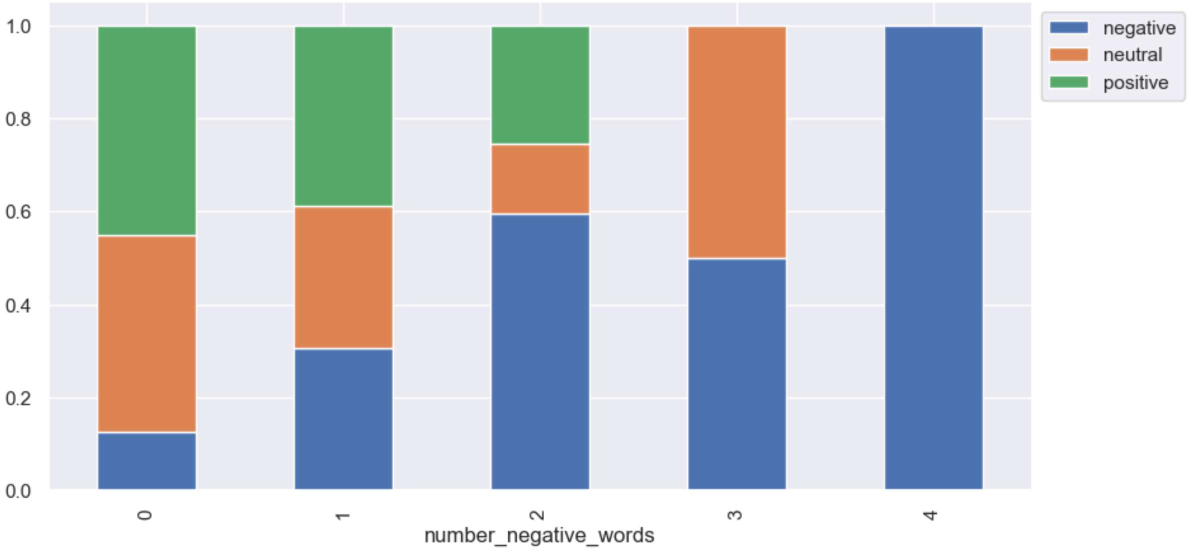


Figure 18 Bar chart presentation of Sentiments as per number of count for Negative Words

Regarding negative words, it can be observed from figures 17 and 18 that tweets with 0 or 1 count for the number of negative words exhibit all three sentiments: 'neutral', 'positive', and 'negative'. On the other hand, tweets with 2, 3, or 4 negative words are predominantly associated with a 'negative' sentiment, as anticipated.

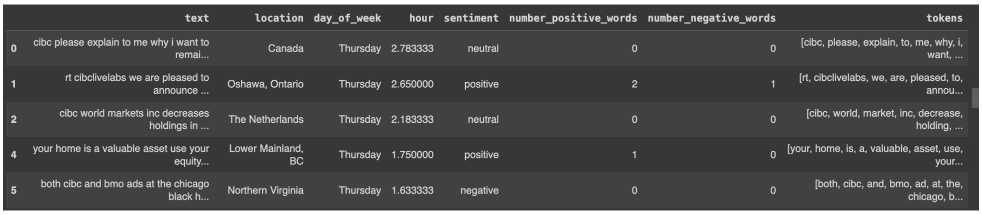
7. 9 Tokenization

Tokenization is the process of slicing up raw text. Tokenization divides unprocessed text into word and phrase tokens. These tokens aid in context comprehension and the development of the NLP model. Sentences, words, characters, and subwords can all be divided by tokenization. Sentence tokenization is the process of dividing text into sentences. It is referred to as word tokenization for words. The tokenize module of NLTK (Natural Language Toolkit), an open source Python library for natural language processing, makes it simple to tokenize textual phrases and words. Tokenize every tweet, the word_tokenize function from nltk.tokenize was imported. Tokenization uses word sequence analysis to help determine the meaning of the text. The following factors make tokenization necessary:

1) Text processing: Tokenization makes it simpler to handle and analyze large volumes of text data by dividing it into smaller, more manageable units.
2) Feature extraction: Tokens are the fundamental building blocks that NLP models use to extract features, which enables them to recognize the underlying patterns and structure in the text.
3) Vocabulary development: Tokenization, where each token represents a unique word or subword, helps a language model's vocabulary grow. For training and inference sequence encoding and decoding, this vocabulary is crucial.

7. 10 Lemmatization

In figure 19, it can be seen the examples of words being broken down to the root like markets has become market & ads has become ad.

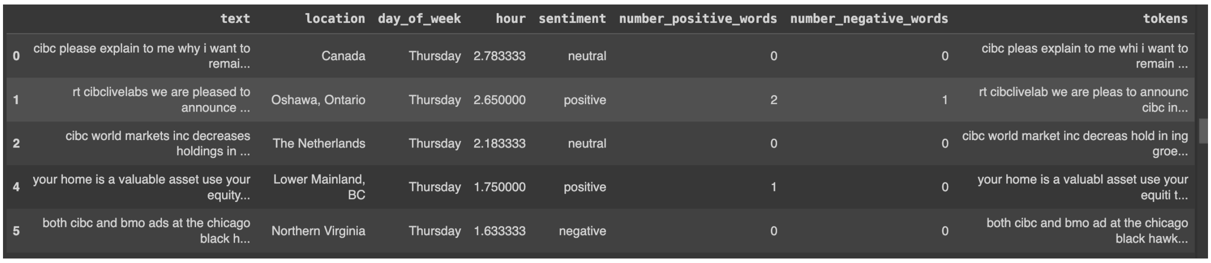


	text	location	day_of_week	hour	sentiment	number_positive_words	number_negative_words	tokens
0	cibc please explain to me why i want to remai...	Canada	Thursday	2.783333	neutral	0	0	[cibc, please, explain, to, me, why, i, want, ...
1	rt cibclivelabs we are pleased to announce ...	Oshawa, Ontario	Thursday	2.650000	positive	2	1	[rt, cibclivelabs, we, are, pleased, to, annou...
2	cibc world markets inc decreases holdings in ...	The Netherlands	Thursday	2.183333	neutral	0	0	[cibc, world, market, inc, decrease, holding, ...
4	your home is a valuable asset use your equity...	Lower Mainland, BC	Thursday	1.750000	positive	1	0	[your, home, is, a, valuable, asset, use, your...
5	both cibc and bmo ads at the chicago black h...	Northern Virginia	Thursday	1.633333	negative	0	0	[both, cibc, and, bmo, ad, at, the, chicago, b...

Figure 19 Lemmatization of words

7. 11 Stemming

Stemming is a crucial component of the natural language processing pipeline. In order to reduce the diversity of tokens in the data, the tokenized words are inputted into the PorterStemmer() function, which converts them to their base form. The words, irrespective of their tense, are visually displayed in the table form in Figure 20 below, with their constituent roots identified. For instance, the words "please" and "pleased" have been reduced to "pleas".

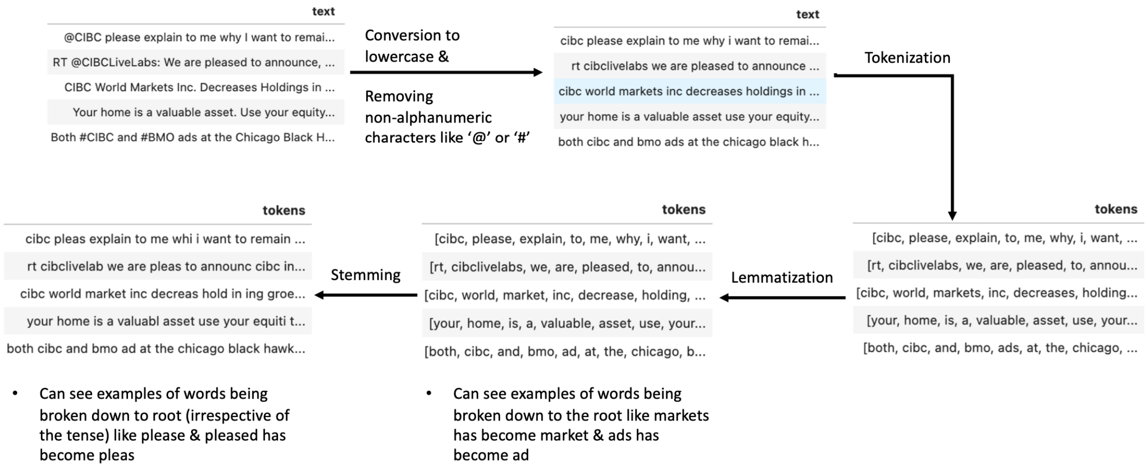


	text	location	day_of_week	hour	sentiment	number_positive_words	number_negative_words	tokens
0	cibc please explain to me why i want to remai...	Canada	Thursday	2.783333	neutral	0	0	cibc pleas explain to me whi i want to remain ...
1	rt cibclivelabs we are pleased to announce ...	Oshawa, Ontario	Thursday	2.650000	positive	2	1	rt cibclivelab we are pleas to announc cibc in...
2	cibc world markets inc decreases holdings in ...	The Netherlands	Thursday	2.183333	neutral	0	0	cibc world market inc decreas hold in ing groe...
4	your home is a valuable asset use your equity...	Lower Mainland, BC	Thursday	1.750000	positive	1	0	your home is a valuabl asset use your equiti t...
5	both cibc and bmo ads at the chicago black h...	Northern Virginia	Thursday	1.633333	negative	0	0	both cibc and bmo ad at the chicago black hawk...

Figure 20 Stemming of words

7. 12 Overall preview of data preprocessing:

Figure 21 provides a summary of the entire data preprocessing process. It begins with converting a raw text into lowercase and removing non-alphanumeric characters, and concludes with the stemming step.



Conversion to lowercase &
Removing non-alphanumeric characters like '@' or '#'
Tokenization
Lemmatization
Stemming

text
@CIBC please explain to me why I want to remai...
RT @CIBCLiveLabs: We are pleased to announce, ...
CIBC World Markets Inc. Decreases Holdings in ...
Your home is a valuable asset. Use your equity...
Both #CIBC and #BMO ads at the Chicago Black H...

text
cibc please explain to me why i want to remai...
rt cibclivelabs we are pleased to announce ...
cibc world markets inc decreases holdings in ...
your home is a valuable asset use your equity...
both cibc and bmo ads at the chicago black h...

tokens
[cibc, please, explain, to, me, why, i, want, ...
[rt, cibclivelabs, we, are, pleased, to, annou...
[cibc, world, markets, inc, decreases, holding...
[your, home, is, a, valuable, asset, use, your...
[both, cibc, and, bmo, ads, at, the, chicago, ...

tokens
[cibc, please, explain, to, me, why, i, want, ...
[rt, cibclivelabs, we, are, pleased, to, annou...
[cibc, world, market, inc, decrease, holding, ...
[your, home, is, a, valuable, asset, use, your...
[both, cibc, and, bmo, ad, at, the, chicago, b...

- Can see examples of words being broken down to the root like markets has become market & ads has become ad

tokens
cibc pleas explain to me whi i want to remain ...
rt cibclivelab we are pleas to announc cibc in...
cibc world market inc decreas hold in ing groe...
your home is a valuabl asset use your equiti t...
both cibc and bmo ad at the chicago black hawk...

- Can see examples of words being broken down to root (irrespective of the tense) like please & pleased has become pleas

Figure 21 Steps of data preprocessing

Following the preprocessing of the data, the process of feature extraction begins. The text is transformed and expressed as vectors to facilitate processing by machine learning algorithms.

7. 13 Feature extraction:

It is best to represent the text as a vector so that machine learning algorithms can process it. Furthermore, it is challenging to learn from large volumes of data in text categorization due to the large number of words, concepts, and phrases. This makes the method computationally expensive. Repetitive and unnecessary features in classification models negatively impact performance and accuracy. Therefore, it is best to extract characteristics in order to reduce the large amount of data and avoid working with high dimensional data. To extract features, utilize CountVectorizer(). Based on the frequency (count) of each word in the text, it turns a given text into a vector. The words are not saved as strings by CountVectorizer(). Rather, they are given an index value. In the matrix produced by CountVectorizer(), each unique token is represented by a column, and each text sample from the data is represented by a row. The value of each cell indicates how many words are contained in that passage of text. The train test split method from sklearn.model selection was used to split the dataset, with 80% of the data used for training and 20% for testing. To ensure that the training data included roughly equal amounts of cyberbullying and non cyberbullying tweets, splitting was done at random. Both the test and train data were converted into their numerical equivalents prior to the model being trained using machine learning or deep learning techniques.

8. Model building

Continuing from the thorough preprocessing and feature extraction stages, the subsequent step in the research involves creating and assessing predictive models. The purpose of this section is to utilize the processed data and extracted features to train different machine learning algorithms. The objective is to determine the most efficient models for predicting cyberbullying incidents. This section aims to provide a comprehensive explanation of the methodologies employed for model training. It will cover the techniques utilized to handle class imbalances, the specific algorithms implemented, and the procedure for hyperparameter tuning. The objective is to guarantee that the models are both precise and resilient, capable of being applied to various datasets. The knowledge acquired from this phase will be vital for comprehending the advantages and disadvantages of each model, ultimately directing the choice of the most effective approach for cyberbullying detection.

8. 1 Label Encoding and One-Hot Encoding

8.1.1 Label Encoding of Target:

Label encoding is a method employed to transform categorical data into numerical data. In this instance, the target variable, which signifies sentiment, is transformed into three numerical values: 1 for 'positive', 0 for 'neutral', and -1 for 'negative'. The process of transformation is crucial for machine learning models to efficiently process the data, as the majority of algorithms necessitate numerical inputs. By utilizing this method of encoding sentiment labels, we can streamline the training process of our models, allowing them to effectively differentiate between various sentiment classes.

8.1.2 One-Hot Encoding of Categorical Columns:

One-hot encoding is used for categorical columns, such as 'day_of_week'. One-hot encoding converts categorical variables into a format suitable for machine learning algorithms, enhancing their predictive capabilities. This entails generating novel binary columns for every distinct category. As an illustration, the 'day_of_week' column would be transformed into seven distinct columns, each corresponding to a different day of the week (such as 'Monday', 'Tuesday', etc.), with binary values denoting the occurrence of that specific day. This method circumvents any hierarchical connection between categories that may be implied by label encoding, guaranteeing that the machine learning models do not make any assumptions about the inherent sequence of the days of the week.

8. 2 TF-IDF Vectorizer for Text Feature Engineering

8.2.1 TF-IDF Vectorizer:

Text feature engineering uses TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. This method weights words according to their frequency and rarity within the dataset, turning the text into a numerical matrix. Words that are specific to a given document are highlighted by the TF-IDF score, which rises with the number of times a word appears in a document but is offset by the number of documents that contain the word. This is especially helpful for finding relevant terms in tweets that might contain crucial sentiment data.

8.2.2 Handling Common English Words:

The TF-IDF vectorizer's `stop_words='english'} parameter is used to eliminate common English words, also known as stop words, such as 'if', 'but', 'or', 'an', and 'the', in order to enhance the quality of the text features. Eliminating these terms contributes to lowering noise and concentrating attention on more significant terms because they typically don't convey any particular emotion. This preprocessing step improves the model's performance by strengthening its capacity to learn from pertinent features.

8. 3 Addressing Target Imbalance with RandomOverSampler

8.3.1 RandomOverSampler:

The dataset's target variable is unbalanced; there are more positive and neutral tweets than negative ones. RandomOverSampler is used to address this. Until the number of examples in each class is roughly balanced, this technique replicates examples from the minority class (negative tweets). By doing this, a more balanced dataset is used to train the machine learning models, which enhances their accuracy and robustness overall and helps them perform better on underrepresented classes.

8. 4 Choice of Machine Learning Algorithms

8.4.1 Popular ML Algorithms for Text Data:

Following a thorough examination of multiple algorithms in the literature review section, four well respected machine learning algorithms were chosen to process text data:

- **Multinomial Naive Bayes:** Well-known for being straightforward and efficient when solving text classification issues.
- **Linear Support Vector Classifier (SVC):** This classifier works well in high dimensional spaces and is useful when there are more dimensions than samples.

- **Random Forest Classifier:** An ensemble technique that averages the outcomes of several decision trees to improve prediction accuracy and reduce overfitting.
- **XGBoost:** An effective gradient boosting framework that performs admirably in terms of speed and accuracy.

These algorithms offer a variety of methods, making it possible to assess each one thoroughly when it comes to how well it performs on the text data.

8. 5 Choice of Metric:

8.5.1 F1 Score:

The principal assessment metric is the F1 score. It provides a balance between recall and precision by being the harmonic mean of the two. Because it guarantees that both false positives and false negatives are taken into account, the F1 score is especially helpful in situations where there is a class imbalance. It seeks to maximize the model's capacity to accurately predict each class by optimizing for the F1 score, guaranteeing a strong performance across all target classes.

8. 6 Machine Learning Workflow

8.6.1 Dataset Splitting:

The dataset is divided in two, 80% for the training set and 20% for the testing set. In order to evaluate the model's performance on data that hasn't been seen yet, this makes sure that it can be trained on one subset and assessed on another.

8.6.2 Cross-Validation:

On the training set, cross-validation is used to find the optimal base vectorizer and machine learning algorithm. To make sure the model generalizes well, this entails splitting the training data into several folds and validating the model on each fold.

8.6.3 Grid Search Cross-Validation:

To fine tune the hyperparameters for the selected vectorizer and ML algorithm, Grid Search Cross-Validation is applied on the training set. Finding the ideal parameters that produce the best model performance is made easier with the help of this thorough search over the given parameter values.

8.6.4 Final Classifier Predictions:

Using the optimal hyperparameters discovered through grid search, the final classifier is applied to the testing set in order to generate predictions. In this step, the model's performance on hypothetical data is assessed, giving an accurate picture of its predictive power.

It guarantees a comprehensive and methodical approach to developing and validating a strong machine learning model for sentiment analysis for an overall comparative analysis by adhering to this structured workflow (illustrated in the Figure 22 below).

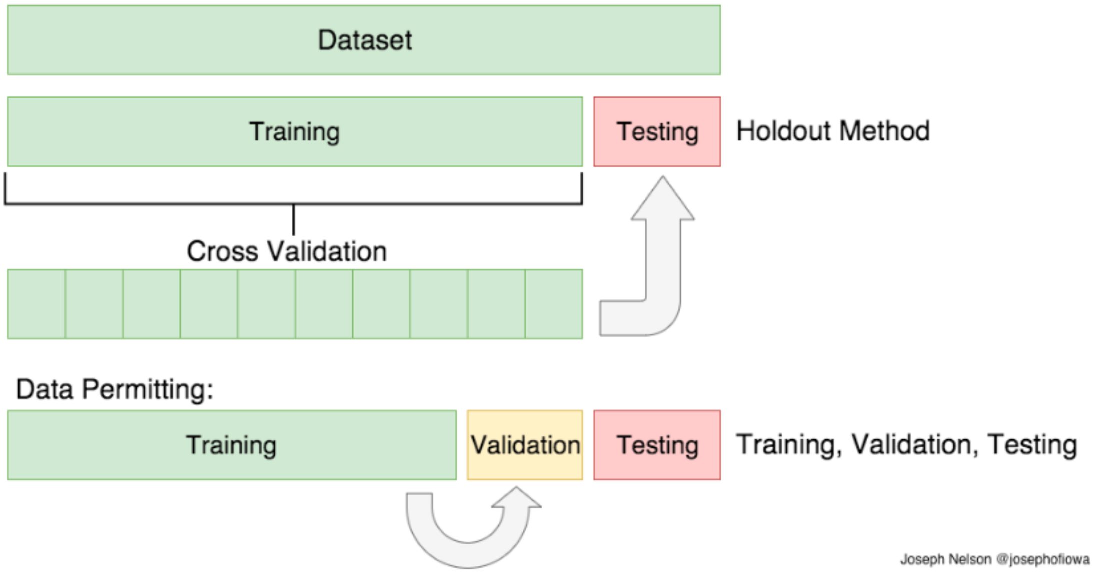


Figure 22 Structured Workflow (Heras & Matovelle, 2021)

9. Results

This chapter presents the findings of this research, focusing on the comparative performance of various machine learning models for cyberbullying detection. The chapter is divided into three main sections; 5-fold Cross Validation Results following with Hyperparameter Tuning with GridSearchCV for Random Forest and then the last section Final Classifier.

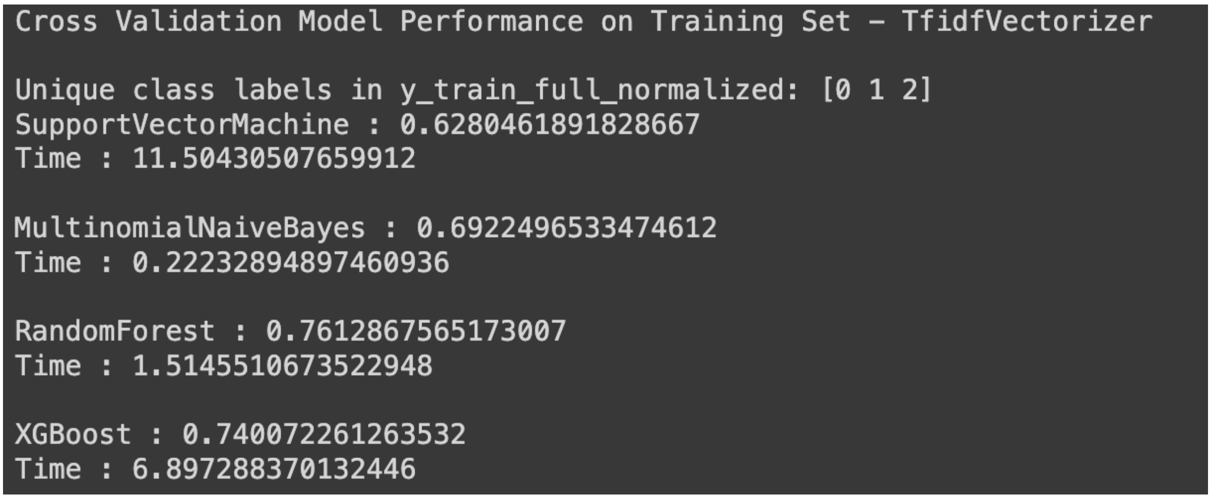
9. 1 5-Fold Cross Validation Results

9.1.1 Evaluation of Models:

During the 5-fold cross-validation process, the performance of various machine learning algorithms was thoroughly assessed to identify the most effective model for the comparative analysis task. This involved splitting the training data into five subsets and iteratively using four subsets for training and one for validation, ensuring that each subset was used for validation once.

The results (shown in figure 23) from this cross-validation indicated that the **Random Forest** and **XGBoost** algorithms achieved the best scores that were 0.761 and 0.740 respectivelyy. The Multinominal Naïve Bayes took the least time to fit and predict while XGBoost and SVM took the maximum amount of time.

Its important to note here that Random Forest is easier to interpret with class_features _importance built in the module with not much compromise in performance score and time.



```
Cross Validation Model Performance on Training Set - TfidfVectorizer

Unique class labels in y_train_full_normalized: [0 1 2]
SupportVectorMachine : 0.6280461891828667
Time : 11.50430507659912

MultinomialNaiveBayes : 0.6922496533474612
Time : 0.22232894897460936

RandomForest : 0.7612867565173007
Time : 1.5145510673522948

XGBoost : 0.740072261263532
Time : 6.897288370132446
```

Figure 23 Cross Validation Model Perfomance on Trainning Set - TfidfVectorizer

The following dicusses each compared algoirthm in detail according to their performance and efficiency.

Multinomial Naive Bayes:

Performance: Multinomial Naive Bayes demonstrated strong performance in classifying the text data.

Efficiency: This algorithm is highly efficient, taking the least amount of time to fit and predict. Its simplicity and speed make it an attractive choice for real time applications and scenarios where quick results are needed.

XGBoost:

Performance: XGBoost also performed exceptionally well, delivering high accuracy and robustness in predictions.

Efficiency: Despite its superior performance, XGBoost requires the second most maximum amount of time for fitting and predicting compared to the other models. This is due to its complex nature and extensive hyperparameter tuning.

Random Forest Classifier:

Interpretability: The Random Forest Classifier, the top performer in terms of raw scores, offers a significant advantage in terms of interpretability. It includes a built-in feature for class feature importance, allowing for a clearer understanding of which features contribute most to the predictions.

Performance and Time: The Random Forest Classifier strikes a balance between performance score and time efficiency. It does not compromise significantly on the performance front and remains relatively time efficient compared to XGBoost, making it a viable option for scenarios where interpretability is crucial.

9.1.2 Summary

In summary, the cross-validation results highlighted the strengths and trade-offs of each algorithm:

- **Multinomial Naive Bayes** is ideal for quick and efficient predictions.
- **XGBoost** excels in performance but at the cost of increased computational time.
- **Random Forest Classifier** provides an easy-to-interpret model with a reasonable balance between performance and time.

By understanding these trade-offs, more informed decisions can be made on selecting the appropriate model based on the specific requirements of the comparative analysis to sentiment analysis task and to detect cyberbullying , whether it prioritizes speed, performance, or interpretability.

The next step will involve tuning hyperparameters for the vectorizer and the chosen model in an effort to improve the performance further.

9. 2 Hyperparameter Tuning with GridSearchCV for Random Forest

9.2.1 Building a Pipeline:

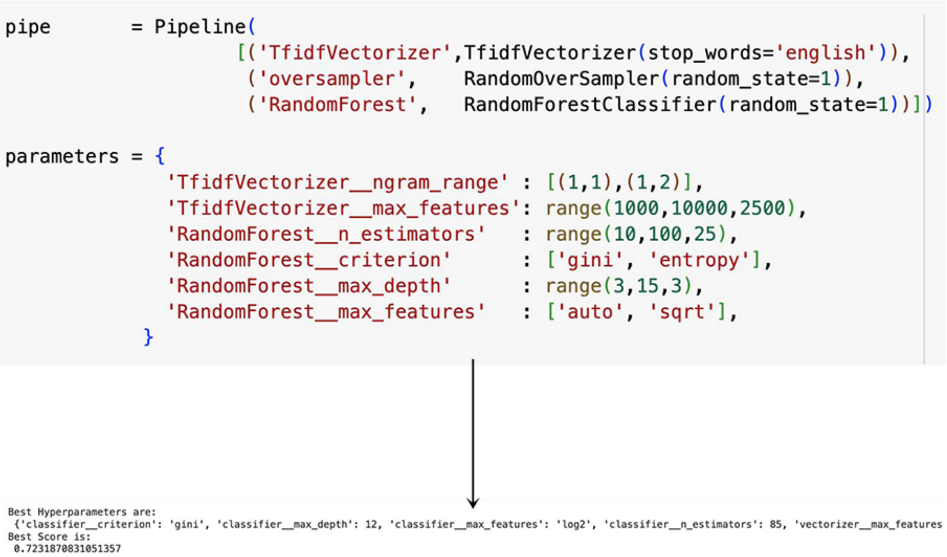
To optimize both the TF-IDF Vectorizer and the Random Forest Classifier, a pipeline in conjunction with GridSearchCV is used. This approach ensures a seamless workflow where each step, from text transformation to model fitting, is executed within a single framework. The pipeline allows to streamline the process of applying transformations and training the model, while GridSearchCV performs an exhaustive search over specified hyperparameter values to identify the best combination.

9.2.2 Steps in the Pipeline:

1. **TF-IDF Vectorizer:** Transform the text data into numerical features.
2. **RandomOverSampler:** Handle class imbalance by oversampling the minority class.
3. **Random Forest Classifier:** Train the model using the transformed features.

9.2.3 Setting Up GridSearchCV:

GridSearchCV will be configured to search over a range of hyperparameters for both the TF-IDF Vectorizer and the Random Forest Classifier. This ensures to find an optimal settings that yield the best performance.



```
pipe       = Pipeline(
                [('TfidfVectorizer',TfidfVectorizer(stop_words='english')),
                 ('oversampler',    RandomOverSampler(random_state=1)),
                 ('RandomForest',   RandomForestClassifier(random_state=1))])

parameters = {
               'TfidfVectorizer__ngram_range' : [(1,1),(1,2)],
               'TfidfVectorizer__max_features': range(1000,10000,2500),
               'RandomForest__n_estimators'   : range(10,100,25),
               'RandomForest__criterion'      : ['gini', 'entropy'],
               'RandomForest__max_depth'      : range(3,15,3),
               'RandomForest__max_features'   : ['auto', 'sqrt'],
             }
```

```
Best Hyperparameters are:
 {'classifier__criterion': 'gini', 'classifier__max_depth': 12, 'classifier__max_features': 'log2', 'classifier__n_estimators': 85, 'vectorizer__max_features
Best Score is:
 0.7231870831051357
```

Figure 24 GridSearch CV Setup

The screenshot taken from compiler in Figure 24 illustrates the ideal arrangement of settings and coding techniques to attain the highest performance score. A detailed explanation of this is provided in the subsequent section.

9.2.4 Explanation:

1. **Pipeline Definition:**
   - The pipeline consists of three main steps: vectorizer for the TF-IDF Vectorizer, oversampler for handling class imbalance, and classifier for the Random Forest Classifier.
2. **Parameter Grid:**
   - The param_grid specifies the exact hyperparameters that were found to be optimal.

- **TF-IDF Vectorizer Parameters:**
  - max_features: Maximum number of features to include.
  - ngram_range: Range of n-values for different n-grams to be extracted (unigrams in this case).
- **Random Forest Classifier Parameters:**
  - n_estimators: Number of trees in the forest.
  - max_depth: Maximum depth of the trees.
  - criterion: Function to measure the quality of a split.
  - max_features: Number of features to consider when looking for the best split.

3. **GridSearchCV Initialization:**
   - GridSearchCV_ is a custom function that fits the pipeline to the training data, searching for the best combination of hyperparameters using StratifiedKFold for cross-validation.
4. **Fitting the Model:**
   - The function GridSearchCV_ iterates over all parameter combinations, performing cross-validation to evaluate each combination's performance and identify the best settings.
5. **Output the Results:**
   - The function prints the best hyperparameters and the best F1 score obtained from the search.

9.2.5 Results:

- **Best Hyperparameters:** {'classifier__criterion': 'gini', 'classifier__max_depth': 12, 'classifier__max_features': 'log2', 'classifier__n_estimators': 85, 'vectorizer__max_features': 1000, 'vectorizer__ngram_range': (1, 1)}
- **Best Score:** 0.7231870831051357

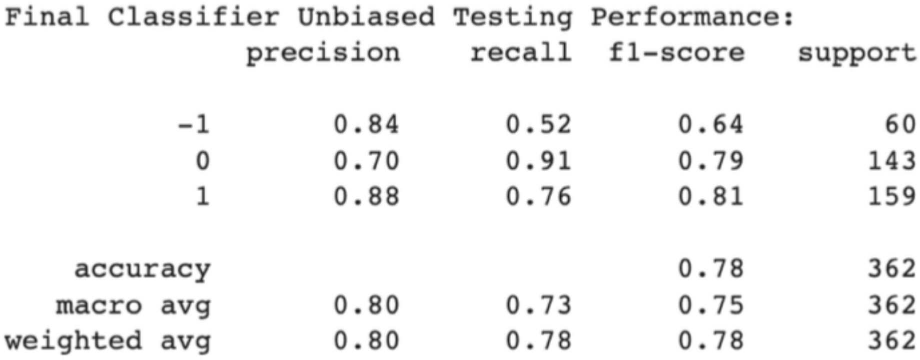
By using this pipeline and GridSearchCV with the specified hyperparameters, both the feature extraction process and the model training are optimized for the best possible performance on the sentiment analysis task. The best F1 score obtained from this tuning process is 0.723, indicating a well performing model.

The last step, the Final Classifier, will involve the enhancement of the performance metrice after tuning following with the indentification of feature impartance of all three classes (positive, negative and neutral) in the final model.

9. 3 Final Classifier

Performance on (unseen) Testing Dataset

The ultimate classifier attained an F1 score of approximately 65% for negative sentiments, and equal to or greater than 75% for positive and neutral sentiments, as depicted in Figure 25.

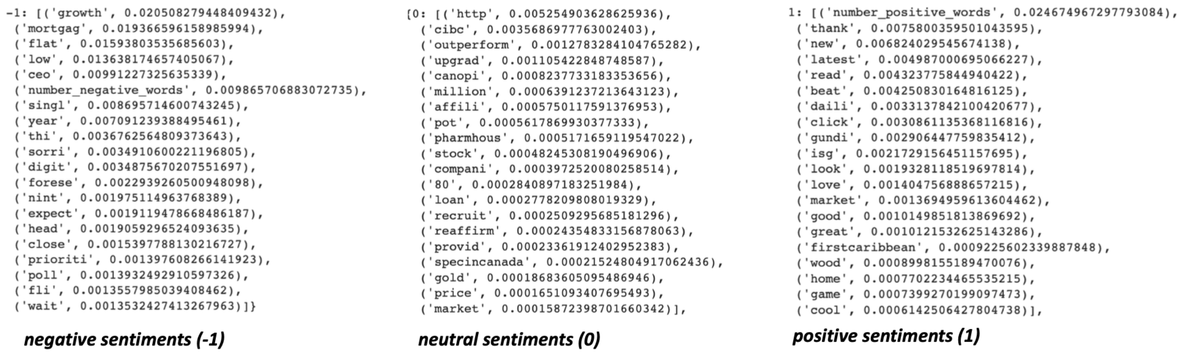


```
Final Classifier Unbiased Testing Performance:
              precision    recall  f1-score   support

          -1       0.84      0.52      0.64        60
           0       0.70      0.91      0.79       143
           1       0.88      0.76      0.81       159

    accuracy                           0.78       362
   macro avg       0.80      0.73      0.75       362
weighted avg       0.80      0.78      0.78       362
```

Figure 25 Scores for the Final Classifier upon Testing Dataset

Feature Importance

Figure 26 illustrates how feature importance was determined for the final model across the three classes of tweets: negative, neutral, and positive. More specifically, "number_positive_words" was critical for predicting positive sentiments and "number_negative_words" had a high importance for predicting negative sentiments. Important terms linked to unfavorable tweets included "flat," "mortgage," "single," "low," "sorry," and "close." Important terms for positive tweets were "new," "thank you," "wood," "latest," "game," "love," "home," "great," and "good." 'http', 'cibc', whereas 'company','recruit','reaffirm', and 'provide' were significant words in neutral tweets. The model's ability to discern sentiment based on pertinent terms is demonstrated by the feature importance words and tweet sentiments that match intuitive expectations.



```
-1: [('growth', 0.020508279448409432),
 ('mortgag', 0.019366596158985994),
 ('flat', 0.01593803535685603),
 ('low', 0.013638174657405067),
 ('ceo', 0.00991227325635339),
 ('number_negative_words', 0.009865706883072735),
 ('singl', 0.008695714600743245),
 ('year', 0.007091239388495461),
 ('thi', 0.003676256480937364 3),
 ('sorri', 0.003491060022119680 5),
 ('digit', 0.003487567020755169 7),
 ('forese', 0.002293926050094809 8),
 ('nint', 0.001975114963768389),
 ('expect', 0.001911947866848618 7),
 ('head', 0.001905929652409363 5),
 ('close', 0.001539778813021672 7),
 ('prioriti', 0.001397608266141923),
 ('poll', 0.001393249291059732 6),
 ('fli', 0.001355798503940846 2),
 ('wait', 0.001353242741326796 3)]}
```

negative sentiments (-1)

```
{0: [('http', 0.005254903628625936),
  ('cibc', 0.0035686977763002403),
  ('outperform', 0.0012783284104765282),
  ('upgrad', 0.001105422848748587),
  ('canopi', 0.0008237733183353656),
  ('million', 0.0006391237213643123),
  ('affili', 0.0005750117591376953),
  ('pot', 0.0005617869930377333),
  ('pharmhous', 0.0005171659119547022),
  ('stock', 0.0004824530819049690 6),
  ('compani', 0.0003972520080258514),
  ('80', 0.0002840897183251984),
  ('loan', 0.0002778209808019329),
  ('recruit', 0.0002509295685181296),
  ('reaffirm', 0.0002435483315687806 3),
  ('provid', 0.0002336191240295238 3),
  ('specincanada', 0.0002152480491706243 6),
  ('gold', 0.0001868360509548694 6),
  ('price', 0.000165109340769549 3),
  ('market', 0.0001587239870166034 2)],
```

neutral sentiments (0)

```
1: [('number_positive_words', 0.024674967297793084),
 ('thank', 0.007580035950104359 5),
 ('new', 0.006824029545674138),
 ('latest', 0.004987000695066227),
 ('read', 0.004323775844940422),
 ('beat', 0.004250830164816125),
 ('daili', 0.003313784210042067 7),
 ('click', 0.003086113536811681 6),
 ('gundi', 0.002906447759835412),
 ('isg', 0.002172915645115769 5),
 ('look', 0.001932811851969781 4),
 ('love', 0.001404756888657215),
 ('market', 0.001369495961360446 2),
 ('good', 0.001014985181386969 2),
 ('great', 0.001012153262514328 6),
 ('firstcaribbean', 0.000922560233988784 8),
 ('wood', 0.000899815518947007 6),
 ('home', 0.000770223446553521 5),
 ('game', 0.000739927019909747 3),
 ('cool', 0.000614250642780473 8)],
```

positive sentiments (1)

Figure 26 Feature Importance

10. Discussion

In this study, comparative analysis was conducted on a dataset of tweets using various natural language processing techniques and machine learning algorithms to detect cyberbullying. To frame this research according to the PSC theory principles, sentiment analysis was crucial in terms of finding out emotions and sentiments attached to each tweet in the chosen dataset. By leveraging TF-IDF vectorization, Random Forest classification, and rigorous hyperparameter tuning, the model achieved robust performance on sentiment classification tasks. The insights gained from feature importance analysis further enriched the understanding of how specific words influence tweet sentiments, highlighting the interpretability and practical applicability of the developed model. These findings contribute to the broader field of cyberbullying detection and demonstrate practical implications for understanding public sentiment through social media data. The following sections summarizes the key findings derived from the research:

10. 1 Comparative Analysis of ML algorithms for Cyberbullying Detection

This section aims to answer the first and main research question " How do different machine learning algorithms (e.g., Logistic Regression, Random Forest, SVM) compare in terms of accuracy, precision, recall, and F1 score for cyberbullying detection? ". The findings of this research demonstrate significant advancments in the detection of cyberbullying on the social media platforms using various machine learnig algorithms. Logistic Regression, XGBoost, Support Vector Classifier machine and Multinomial Naïve Bayes were the four models that were compared on upon roughly 2000 scrapped tweets. This dataset was taken from University of California computer science department as mentioned above in data description section. Their comparative anaylsis provided an insightful results regarding their respective performances.

10.1.1 Perfomances of Machine Learning Algorithms

- **Multinomial Naïve Bayes:** This model excelled in scenarios requiring quick predictions due to its efficient computation. Its simplicity and speed make it a strong candidate for real time cyberbullying detection systems.
- **Random Forest:** This model offered a robust balance between performance and interpretability. It showed high accuracy and provided meaningful insights into the feature importance, which is valuable for understanding the factors contributing to cyberbullying.
- **XGBoost:** This model achieved superior accuracy but at the cost of increased computational resources. Its complexity and resource demands may limit its application in real time scenarios but highlight its potential for batch processing where accuracy is paramount.
- **Support Vector Machine:** This model took the maximum amount of time to compute and its result scores were also the lowest in comparison to the others.

All the models were compared through 5-fold Cross validation process. The cross-validation results (presented in Table X) demonstrated that the Random Forest and XGBoost algorithms yielded the highest scores, with 0.761 and 0.740, respectively. XGBoost and SVM required the most time to fit and predict, whereas Multinominal Naïve Bayes required the least time.
It is noteworthy to mention that Random Forest's class_features_importance module makes it easier to interpret while maintaining a high performance score and speed.

10. 2 Influence of Sentiment Analysis Features on Predictive Models

This section aims to answer the sub research question ‘a’: How does the inclusion of sentiment analysis features influence the effectiveness of predictive models in identifying cyberbullying?

The models' ability to predict the instances of cyberbullying was greatly improved by the addition of sentiment analysis features. The significance of emotional context in identifying harmful online behavior is highlighted by this finding. TextBlob was used to categorize the tweets into three sentiment categories: negative, neutral, and positive. With sentiment ratios of 0.15 for negative, 0.43 for neutral, and 0.40 for positive tweets, the dataset showed an inherent class imbalance. For that the dataset was analysed upon feature importance as dicussed below.

10.2.1 Feature Importance Analysis

Feature importance analysis was done after the model was trained and evaluated to determine which features were most important in predicting each sentiment class (positive, neutral, and negative). It is noteworthy that certain features, like {number_negative_words}, were crucial in predicting negative sentiments, whereas `number_positive_words} was crucial in predicting positive sentiments.

10.2.2 Key Words Associated with Sentiments

The study identified specific words that were crucial for each sentiment class:

- **Negative Sentiments:** Words like 'flat', 'mortgage', 'single', 'low', 'sorry', and 'close'.
- **Positive Sentiments:** Words like 'new', 'thank', 'wood', 'latest', 'game', 'love', 'home', 'great', and 'good'.
- **Neutral Sentiments:** Words such as 'http', 'cibc', 'company', 'recruit', 'reaffirm', and 'provide'.

Each sentiment class's key words make sense when compared to the sentiment categories they belong to. This shows that the model was successful in identifying associations and patterns in the tweet data related to semantics.

10. 3 Integration of PSC Principles in Predictive Analytics

This section aims to answer the sub research question ‘b’ : “ How can principles of Psychosocial Safety Climate (PSC) be integrated into predictive analytics models to enhance cyberbullying detection? ”

A key component of this research was the incorporation of Psychosocial Safety Climate (PSC) principles into the predictive analytics models. The PSC theory suggests that by fostering a safe and supportive atmosphere, a positive safety climate can lessen undesirable behaviors like bullying.

10.3.1 PSC and Cyberbullying Detection

The arguments were based on the idea that by encouraging a safer social environment, a robust PSC would lessen the incidence of cyberbullying. This assumption is supported by the research results, which demonstrate that characteristics indicating a favorable safety climate lead to more precise forecasts. PSC mechanisms, such as established procedures (such as sentiment scores) for addressing bullying, were represented within the features set like sentimental values and were found to influence the capability of the predictive model positively.

10.3.2 Model Integration and Performance

The models' operationalization of PSC principles was achieved by incorporating features that were obtained from sentiment and linguistic analysis. Assuming that these characteristics will essentially improve the model's performance because they represented the psychosocial setting.
The application of the theoretical framework in predictive analytics for cyberbullying detection was validated in terms of positive resutls in terms of accuracy and processing time that resulted from the incorporation of PSC related features.

10.3.3 Theoratical Implications

Furthermore, the results of this study add to the body of knowledge already available on PSC and the identification of cyberbullying by providing fresh perspectives on the interactions between technological advancements and psychosocial variables.

10.3.4 Advancement of PSC Theory

The findings lend credence to the theory that a positive PSC can reduce cyberbullying by encouraging a safer online community. The increased predictive accuracy that results from incorporating PSC principles into the models serves as evidence of this.
The findings correspond with existing literature indicating that PSC indirectly lowers poor psychological health outcomes by implementing bullying mitigation techniques that are successful.

10. 4 Limitations and recommendations

10.4.1 Limitations

While the research offers valuable contributions, several limitations must be acknowledged

1. **Data Limitation:** It's possible that the dataset utilized for this study does not accurately reflect the variety of cyberbullying incidents that occur on various social media platforms. To improve the generalizability of the results, more diverse datasets should be included in future research.

2. **Model Generalizability:** Different datasets or real world scenarios may yield different results from the models in terms of performance. Furthermore, the practical implementation of some models, such as XGBoost, in real time

applications may be restricted due to the computational resources needed for them.

3. **PSC-related Features:** To align PSC principles completely with the predictive models, along with sentiment score, management commitment index and psychological health metrics are also very important to integrate. These two were the limitations of this research.

10.4.2 Recommendations

Based on the above limitations and findings of this research, several areas for the future investigation are suggested.

1. **Broader Data Collection:** To increase the generalizability of the findings, future research should concentrate on gathering more diverse datasets that cover a broad range of social media platforms and user demographics.

2. **Enhanced Feature Integration:** Cyberbullying detection could significantly advance by investigating other features that could further improve model accuracy, such as incorporating social network analysis or more nuanced sentiment analysis.

3. **Real world Implementation:** Practical application of this research depends on examining the obstacles to and solutions for putting these models into practice in real world scenarios, including addressing privacy concerns and ethical issues.

11. Conclusion

To summarize, this comparative analysis highlights the advantages and disadvantages of different predictive analytics models in forecasting cyberbullying. Both Multinomial Naive Bayes and Random Forest emerged as leading contenders, with each demonstrating excellence in different areas. The study's findings make significant contributions to the field of cyberbullying detection. Additionally, they offer a potential framework for stakeholders, such as moderators and administrators, to guide future research and development in predictive analytics for social media content. It is important to note that social media platforms should also implement policies to effectively address cyberbullying. Utilizing the integrated functionalities on these platforms to report incidents of cyberbullying may not always yield desired outcomes, but it can result in the suspension or limitation of persistently abusive accounts.

The Multinomial Naive Bayes algorithm has demonstrated its suitability for situations that demand fast and effective predictions. The simplicity and speed of this technology make it highly appropriate for real time applications, such as monitoring social media platforms in real time. This is especially important for quickly identifying instances of cyberbullying. The performance of this model demonstrates its ability to efficiently handle text classification tasks using a direct probabilistic approach.

XGBoost, although the most computationally intensive, demonstrated superior performance in terms of accuracy and robustness. Its extensive hyperparameter tuning allows for fine grained control over the model's performance, making it an excellent choice for scenarios where the highest possible accuracy is paramount, and computational resources are abundant. This model is well suited for detailed analyses and offline processing tasks where time constraints are less stringent.

Random Forest, on the other hand, offered a significant advantage in terms of interpretability. Its ability to provide feature importance scores directly contributes to understanding the factors influencing predictions, which is crucial for explainable AI. This model strikes a balance between performance and interpretability, making it a viable option for applications where understanding the decision making process is as important as the prediction itself. This feature is particularly valuable in contexts such as academic research, policymaking, and developing transparent AI systems where stakeholders need to comprehend the underlying mechanics of the model.

The feature importance analysis provided valuable insights into the types of words associated with different sentiments, aiding in the development of more targeted intervention strategies. For instance, identifying words indicative of negative sentiment can help moderators focus on potentially harmful content, while recognizing positive engagement terms can enhance the identification of supportive and constructive interactions.

The approach of addressing data imbalance through techniques such as RandomOverSampler underscores the importance of ensuring fair and accurate predictions across all sentiment classes. By balancing the training dataset, the models were able to perform more effectively, particularly in predicting minority classes like negative sentiment, which is often underrepresented in datasets.

By leveraging these insights, future research can build on the strengths of these models to develop more sophisticated and nuanced approaches to cyberbullying detection. The study provides a clear framework for selecting appropriate models based on specific requirements, whether prioritizing speed, accuracy, or interpretability. Moreover, it lays the groundwork for integrating these models into comprehensive systems aimed at enhancing the safety and wellbeing of online communities. Implementing such systems can lead to more effective monitoring and moderation of social media content, ultimately contributing to a safer online environment where individuals can interact without fear of harassment or bullying.

12. Bibliography

Ahlfors, R. (2010). Many Sources, One Theme: Analysis of Cyberbullying Prevention and Intervention Websites. *Journal of Social Sciences*, *6*(4), 515–522. https://doi.org/10.3844/jssp.2010.515.522

Akhter, A., Acharjee, U., & Polash, M. (2019). Cyber Bullying Detection and Classification using Multinomial Naïve Bayes and Fuzzy Logic. In *International Journal of Mathematical Sciences and Computing* (Vol. 5). https://doi.org/10.5815/ijmsc.2019.04.01

Albert Bandura's Social Learning Theory In Psychology. (2024, February 1). https://www.simplypsychology.org/bandura.html

Al-Garadi, M., Hussain, M., Khan, N., Murtaza, G., Nweke, H., Ihsan, A., Mujtaba, G., Chiroma, H., Khattak, H. A., & Gani, A. (2019). Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges. *IEEE Access*, *PP*, 1–1. https://doi.org/10.1109/ACCESS.2019.2918354

Al-Garadi, M., Varathan, K., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, *63*, 433–443. https://doi.org/10.1016/j.chb.2016.05.051

Anderson, C., & Bushman, B. (2002). Human Aggression. *Annual Review of Psychology*, *53*, 27–51. https://doi.org/10.1146/annurev.psych.53.100901.135231

Ang, R., & Goh, D. (2010). Cyberbullying Among Adolescents: The Role of Affective and Cognitive Empathy, and Gender. *Child Psychiatry and Human Development*, *41*, 387–397. https://doi.org/10.1007/s10578-010-0176-3

Ates, E. C., Bostanci, E., & Guzel, M. S. (2021). *Comparative Performance of Machine Learning Algorithms in Cyberbullying Detection: Using Turkish Language Preprocessing Techniques* (arXiv:2101.12718). arXiv. https://doi.org/10.48550/arXiv.2101.12718

Baillien, E., Bollen, K., & De Witte, H. (2013). Conflicts and conflict management styles as precursors of workplace bullying: A two-wave longitudinal study. *International Journal of Work Organisation and Emotion*, *23*. https://doi.org/10.1080/1359432X.2012.752899

Barlett, C. (2024). Social learning of cyberbullying perpetration: The interactive role of parent and peer cyberbullying and cyberbullying reinforcement in a sample of U.S. adolescents: A brief report. *Technology, Mind, and Behavior*, *5*. https://doi.org/10.1037/tmb0000126

Barlińska, J., Szuster, A., & Winiewski, M. (2013). Cyberbullying among Adolescent Bystanders: Role of the Communication Medium, Form of Violence, and Empathy. *Journal of Community & Applied Social Psychology*, *Volume 23*, 37–51. https://doi.org/10.1002/casp.2137

Bauman, S., Cross, D., & Walker, J. (Eds.). (2012). *Principles of Cyberbullying Research: Definitions, Measures, and Methodology*. Routledge. https://doi.org/10.4324/9780203084601

Bond, S. A., Tuckey, M., & Dollard, M. (2010). Psychosocial safety climate, workplace bullying, and symptoms of posttraumatic stress. *Organization Development Journal*, *28*, 37–56.

Brownlee, J. (n.d.). *Master Machine Learning Algorithms*.

Buczak, A., & Guven, E. (2015). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*, *18*, 1–1. https://doi.org/10.1109/COMST.2015.2494502

Chatterjee, R., Datta, A., & Sanyal, D. (2018). *Ensemble Learning Approach to Motor Imagery EEG Signal Classification* (pp. 183–208). https://doi.org/10.1016/B978-0-12-816086-2.00008-4

Chatzakou, D., Leontiadis, I., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A., & Kourtellis, N. (2019). Detecting Cyberbullying and Cyberaggression in Social Media. *ACM Transactions on the Web*, *13*, 1–51. https://doi.org/10.1145/3343484

Chavan, V., & S S, S. (2015). *Machine learning approach for detection of cyber-aggressive comments by peers on social media network*. 2354–2358. https://doi.org/10.1109/ICACCI.2015.7275970

Chawla, N. (2005). Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook, ISBN 978-0-387-09822-7. Springer Science+Business Media, LLC, 2010, p. 875* (Vol. 5, pp. 853–867). https://doi.org/10.1007/0-387-25465-X_40

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). *Detecting Offensive Language in Social Media to Protect Adolescent Online Safety*. 71–80. https://doi.org/10.1109/SocialCom-PASSAT.2012.55

Chengsheng, T., Huacheng, L., & Bing, X. (2017). AdaBoost typical Algorithm and its application research. *MATEC Web of Conferences*, *139*, 00222. https://doi.org/10.1051/matecconf/201713900222

Cheyne, A., Cox, S., Oliver, A., & Tomás, J. M. (1998). Modelling safety climate in the prediction of levels of safety activity. *Work & Stress*, *12*(3), 255–271. https://doi.org/10.1080/02678379808256865

Christian, M. S., Bradley, J. C., Wallace, J. C., & Burke, M. J. (2009). Workplace safety: A meta-analysis of the roles of person and situation factors. *Journal of Applied Psychology*, *94*(5), 1103–1127. https://doi.org/10.1037/a0016172

Comparison between Naïve Bayes and Logistic Regression – DataEspresso. (2017, October 24). https://dataespresso.com/en/2017/10/24/comparison-between-naive-bayes-and-logistic-regression/

Cooper-Thomas, H., Gardner, D., O'Driscoll, M., Catley, B., Bentley, T., & Trenberth, L. (2013). Neutralizing workplace bullying: The buffering effects of contextual factors. *Journal of Managerial Psychology*, *28*(4), 384–407. https://doi.org/10.1108/JMP-12-2012-0399

Crawford, W. T., & Detar, W. J. (2023). The Relationship between Job Demands, Job Resources, Employee Burnout, and Employee Engagement in Municipal Government Workers. *Journal of Service Science and Management*, *16*(4), Article 4. https://doi.org/10.4236/jssm.2023.164024

Cyberbullying On Social Media. (2019, October 10). https://helpyourteennow.com/cyberbullying-on-social-media/

Cyberbullying: Where are we now? A cross-national understanding (Printed edition of the special issue published in Societies). | Request PDF. (n.d.). Retrieved April

14, 2024, from https://www.researchgate.net/publication/323735244_Cyberbullying_Where_are_we_now_A_cross-national_understanding_Printed_edition_of_the_special_issue_published_in_Societies

Dadvar, M., de Jong, F., Ordelman, R., & Trieschnigg, D. (2012, January 1). *Improved Cyberbullying Detection Using Gender Information*. Cognitive Processing - COGN PROCESS.

Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). *Improving Cyberbullying Detection with User Context*. pp 693-696. https://doi.org/10.1007/978-3-642-36973-5_62

Deb, S. (2016, March 21). Naive Bayes vs Logistic Regression. *Medium*. https://medium.com/@sangha_deb/naive-bayes-vs-logistic-regression-a319b07a5d4c

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. *ACM Transactions on Interactive Intelligent Systems*, *2*. https://doi.org/10.1145/2362394.2362400

Dinakar, K., Reichart, R., & Lieberman, H. (2011, January 1). *Modeling the Detection of Textual Cyberbullying*.

Dollard, M. F., & Bakker, A. B. (2010). Psychosocial safety climate as a precursor to conducive work environments, psychological health problems, and employee engagement. *Journal of Occupational and Organizational Psychology*, *83*(3), 579–599. https://doi.org/10.1348/096317909X470690

Dollard, M., & Mcternan, W. (2011). Psychosocial safety climate: A multilevel theory of work stress in the health and community service sector. *Epidemiology and Psychiatric Sciences*, *20*, 287–293. https://doi.org/10.1017/S2045796011000588

Einarsen, S., Skogstad, A., Rørvik, E., Lande, Å. B., & Nielsen, M. B. (2018). Climate for conflict management, exposure to workplace bullying and work engagement: A moderated mediation analysis. *The International Journal of Human Resource Management*, *29*(3), 549–570. https://doi.org/10.1080/09585192.2016.1164216

Escartín Solanelles, J., Ceja, L., Navarro, J., & Zapf, D. (2013). Modeling workplace bullying using catastrophe theory. *Nonlinear Dynamics, Psychology, and Life Sciences*, *17*, 493–515.

Fanti, K., Demetriou, A., & Hawa, V. (2012). A longitudinal study of cyberbullying: Examining riskand protective factors. *European Journal of Developmental Psychology*, *9*, 168–181. https://doi.org/10.1080/17405629.2011.643169

Flood, Ed. D., Francesca. (2017). *Social Psychology of Organizations*. https://doi.org/10.1007/978-3-319-31816-5_3059-1

Galán-García, P., Puerta, J., Laorden, C., Santos, I., & Bringas, P. (2014). Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying. In *Advances in Intelligent Systems and Computing* (Vol. 239, pp. 419–428). https://doi.org/10.1007/978-3-319-01854-6_43

García-Recuero, Á. (2016). *Discouraging Abusive Behavior in Privacy-Preserving Online Social Networking Applications*. 305–309. https://doi.org/10.1145/2872518.2888600

Garrick, A., Mak, A., Cathcart, S., Winwood, P., & Lushington, K. (2014). Psychosocial safety climate moderating the effects of daily job demands and recovery on fatigue and work engagement. *Journal of Occupational and Organizational Psychology*, *87*. https://doi.org/10.1111/joop.12069

Goerzig, A., & Ólafsson, K. (2012). What Makes a Bully a Cyberbully? Unravelling the Characteristics of Cyberbullies across Twenty-Five European Countries. *Journal of Children and Media*, *7*, 1–19. https://doi.org/10.1080/17482798.2012.739756

Goldberg, Y., & Levy, O. (2014). *word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method*.

Hall, G., Dollard, M., & Coward, J. (2010). Psychosocial Safety Climate: Development of the PSC-12. *International Journal of Stress Management*, *17*, 353–383. https://doi.org/10.1037/a0021320

Harris, K. J., Harvey, P., Harris, R. B., & Cast, M. (2013). An Investigation of Abusive Supervision, Vicarious Abusive Supervision, and Their Joint Impacts. *The Journal of Social Psychology*, *153*(1), 38–50. https://doi.org/10.1080/00224545.2012.703709

Harryzhang. (2011). EXPLORING CONDITIONS FOR THE OPTIMALITY OF NAÏVE BAYES. *International Journal of Pattern Recognition and Artificial Intelligence*, *19*. https://doi.org/10.1142/S0218001405003983

Hasan, Md. T., Al, E., Hossain, Md. S., Akter, A., Ahmed, M., & Islam, S. (2023). *A Review on Deep-Learning-Based Cyberbullying Detection*. *15*. https://doi.org/10.3390/fi15050179

Hassan Yousef, A., Medhat, W., & Mohamed, H. (2014). Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Engineering Journal*, *5*. https://doi.org/10.1016/j.asej.2014.04.011

Havas, J., Nooijer, J., Crutzen, R., & Feron, F. (2011). Adolescents' views about an Internet platform for adolescents with mental health problems. *Health Education*, *111*, 164–176. https://doi.org/10.1108/09654281111123466

Heras, D., & Matovelle, C. (2021). Machine-learning methods for hydrological imputation data: Analysis of the goodness of fit of the model in hydrographic systems of the Pacific - Ecuador. *Ambiente e Agua - An Interdisciplinary Journal of Applied Science*, *16*, 1. https://doi.org/10.4136/ambi-agua.2708

Hinduja, S., & Patchin, J. (2008). Cyberbullying: An Exploratory Analysis of Factors Related to Offending and Victimization. *Deviant Behavior - DEVIANT BEHAV*, *29*, 129–156. https://doi.org/10.1080/01639620701457816

Idris, M. A., Dollard, M. F., Coward, J., & Dormann, C. (2012). Psychosocial safety climate: Conceptual distinctiveness and effect on job demands and worker psychological health. *Safety Science*, *50*(1), 19–28. https://doi.org/10.1016/j.ssci.2011.06.005

Irena, B., & Setiawan, E. (2020). Fake News (Hoax) Identification on Social Media Twitter using Decision Tree C4.5 Method. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*, 711–716. https://doi.org/10.29207/resti.v4i4.2125

Jacobs, N. C., Völlink, T., Dehue, F., & Lechner, L. (2014). Online Pestkoppenstoppen: Systematic and theory-based development of a web-based tailored intervention for adolescent cyberbully victims to combat and prevent cyberbullying. *BMC Public Health*, *14*(1), 396. https://doi.org/10.1186/1471-2458-14-396

Jiang, C., Zhang, H., Ren, Y., Han, Z., Chen, K.-C., & Hanzo, L. (2016). Machine Learning Paradigms for Next-Generation Wireless Networks. *IEEE Wireless Communications*, *PP*. https://doi.org/10.1109/MWC.2016.1500356WC

Joachims, T. (1998). Text Categorization with Support Vector Machines. *Proc. European Conf. Machine Learning (ECML'98)*. https://doi.org/10.17877/DE290R-5097

Jr, D. W. H., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

KiVa is an anti-bullying programme | KiVa Antibullying Program | Just another KiVa Koulu site. (n.d.). KiVa Program. Retrieved April 16, 2024, from https://www.kivaprogram.net/

Kompier, M., & Kristensen, T. (2000). Organizational work stress interventions in a theoretical, methodological and practical context. *Stress in the Workplace: Past, Present and Future*.

Kwan, S. S. M., Tuckey, M. R., & Dollard, M. F. (2016). The role of the psychosocial safety climate in coping with workplace bullying: A grounded theory and sequential tree analysis. *European Journal of Work and Organizational Psychology*, *25*(1), 133–148. https://doi.org/10.1080/1359432X.2014.982102

Larson, S. (2024, March 26). Social Media Users 2024 (Global Data & Statistics). *Priori Data*. https://prioridata.com/data/social-media-usage/

Law, R., Dollard, M. F., Tuckey, M. R., & Dormann, C. (2011). Psychosocial safety climate as a lead indicator of workplace bullying and harassment, job resources, psychological health and employee engagement. *Accident Analysis & Prevention*, *43*(5), 1782–1793. https://doi.org/10.1016/j.aap.2011.04.010

Learn to Recognize the Real-Life Effects of Cyberbullying on Children. (n.d.). Parents. Retrieved July 17, 2024, from https://www.parents.com/what-are-the-effects-of-cyberbullying-460558

Lenhart, A., Purcell, K., Smith, A., & Zickuhr, K. (2010). Social Media & Mobile Internet Use Among Teens and Young Adults. *Pew Internet and American Life Project*.

LI, J., HUANG, G., FAN, C., SUN, Z., & ZHU, H. (2019). Key word extraction for short text via word2vec, doc2vec, and textrank. *TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES*, *27*, 1794–1805. https://doi.org/10.3906/elk-1806-38

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice*. https://doi.org/10.13140/2.1.1570.5928

Lu, Y., Wang, J., Liu, M., Zhang, K., Gui, G., Ohtsuki, T., & Adachi, F. (2020). Semi-Supervised Machine Learning Aided Anomaly Detection Method in Cellular Networks. *IEEE Transactions on Vehicular Technology*, *PP*, 1–1. https://doi.org/10.1109/TVT.2020.2995160

Lutte contre le harcèlement à l'école. (n.d.). Ministère de l'Education Nationale et de la Jeunesse. Retrieved April 16, 2024, from https://www.education.gouv.fr/non-au-harcelement

Mandot, P. (2018, December 1). What is LightGBM, How to implement it? How to fine tune the parameters? *Medium*. https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc

Mangaonkar, A., Hayrapetian, A., & Raje, R. (2015). *Collaborative detection of cyberbullying behavior in Twitter data*. 611–616. https://doi.org/10.1109/EIT.2015.7293405

Mccallum, A., & Nigam, K. (2001). A Comparison of Event Models for Naive Bayes Text Classification. *Work Learn Text Categ*, *752*.

Misra, S., & Li, H. (2020). *Noninvasive fracture characterization based on the classification of sonic wave travel times* (pp. 243–287). https://doi.org/10.1016/B978-0-12-817736-5.00009-0

Moreno-Jiménez, B., Rodríguez-Muñoz, A., Pastor, J. C., Sanz-Vergel, A. I., & Garrosa, E. (2009). The moderating effects of psychological detachment and thoughts of revenge in workplace bullying. *Personality and Individual Differences*, *46*(3), 359–364. https://doi.org/10.1016/j.paid.2008.10.031

Muneer, A., & Fati, S. (2020a). A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. *Future Internet*, *12*, 187. https://doi.org/10.3390/fi12110187

Muneer, A., & Fati, S. M. (2020b). A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. *Future Internet*, *12*(11), Article 11. https://doi.org/10.3390/fi12110187

Nahar, V., Li, X., Zhang, H. L., & Pang, C. (2014). Detecting cyberbullying in social networks using multi-agent system. *Web Intelligence and Agent Systems*, *12*, 375–388. https://doi.org/10.3233/WIA-140301

Nandakumar, V. (2018). CYBERBULLYING REVELATION IN TWITTER DATA USING NAÏVE BAYES CLASSIFIER ALGORITHM. *International Journal of Advanced Research in Computer Science*, *9*, 510–513. https://doi.org/10.26483/ijarcs.v9i1.5396

Notar, C., Padgett, S., & Roden, J. (2013). Cyberbullying: Resources for Intervention and Prevention. *Universal Journal of Educational Research*, *1*, 133–145. https://doi.org/10.13189/ujer.2013.010301

Novalita, N., Herdiani, A., Lukmana, I., & Puspandari, D. (2019). Cyberbullying identification on twitter using random forest classifier. *Journal of Physics: Conference Series*, *1192*, 012029. https://doi.org/10.1088/1742-6596/1192/1/012029

Oliverio, A. (2023). Social Action Explanation and Intentions in Sociology and Social Sciences. *Advances in Applied Sociology*, *13*(08), 573–586. https://doi.org/10.4236/aasoci.2023.138036

Olweus, D. (2012). Cyberbullying: An overrated phenomenon? *European Journal of Developmental Psychology - EUR J DEV PSYCHOL*, *9*, 1–19. https://doi.org/10.1080/17405629.2012.682358

Online Pestkoppenstoppen: Systematic and theory-based development of a web-based tailored intervention for adolescent cyberbully victims to combat and prevent cyberbullying | BMC Public Health | Full Text. (n.d.). Retrieved April 16, 2024, from https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-14-396

Owen, M., Bailey, T., & Dollard, M. (2016). *Psychosocial Safety Climate as a Multilevel Extension of ERI Theory: Evidence from Australia* (pp. 189–217). https://doi.org/10.1007/978-3-319-32937-6_9

Pan, B. (2018). Application of XGBoost algorithm in hourly PM2.5 concentration prediction. *IOP Conference Series: Earth and Environmental Science*, *113*, 012127. https://doi.org/10.1088/1755-1315/113/1/012127

Patchin, J., & Hinduja, S. (2010). Traditional and Nontraditional Bullying Among Youth: A Test of General Strain Theory. *Youth & Society - YOUTH SOC*, *41*. https://doi.org/10.1177/0044118X10366951

Patchin, J. W. (2024, February 16). *2023 Cyberbullying Data*. Cyberbullying Research Center. https://cyberbullying.org/2023-cyberbullying-data

Patel, S. (2017, May 18). Chapter 5: Random Forest Classifier. *Machine Learning 101*. https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1

Pawar, R., Agrawal, Y., Joshi, A., Gorrepati, R., & Raje, R. (2018). *Cyberbullying Detection System with Multiple Server Configurations*. 0090–0095. https://doi.org/10.1109/EIT.2018.8500110

(PDF) Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. (n.d.).

Retrieved April 16, 2024, from https://www.researchgate.net/publication/331821059_Anatomy_of_Online_Hate_Developing_a_Taxonomy_and_Machine_Learning_Models_for_Identifying_and_Classifying_Hate_in_Online_News_Media

(PDF) AUTOMATIC DETECTION OF CYBERBULLYING IN FORMSPRING.ME, MYSPACE AND YOUTUBE SOCIAL NETWORKS. (n.d.). Retrieved April 14, 2024, from https://www.researchgate.net/publication/336185254_AUTOMATIC_DETECTION_OF_CYBERBULLYING_IN_FORMSPRINGME_MYSPACE_AND_YOUTUBE_SOCIAL_NETWORKS

(PDF) Brute Force Works Best Against Bullying. (n.d.). Retrieved April 16, 2024, from https://www.researchgate.net/publication/280529227_Brute_Force_Works_Best_Against_Bullying

(PDF) Cyberbullying in Children and Youth: Implications for Health and Clinical Practice. (n.d.). Retrieved April 14, 2024, from https://www.researchgate.net/publication/312657450_Cyberbullying_in_Children_and_Youth_Implications_for_Health_and_Clinical_Practice

(PDF) Detection of Cyberbullying Incidents on the Instagram Social Network. (n.d.). Retrieved April 16, 2024, from https://www.researchgate.net/publication/273640275_Detection_of_Cyberbullying_Incidents_on_the_Instagram_Social_Network

(PDF) Fake News (Hoax) Identification on Social Media Twitter using Decision Tree C4.5 Method. (n.d.). Retrieved April 16, 2024, from

https://www.researchgate.net/publication/343941072_Fake_News_Hoax_Identification_on_Social_Media_Twitter_using_Decision_Tree_C45_Method

(PDF) Logistic regression in data analysis: An overview. (n.d.). Retrieved April 14, 2024, from https://www.researchgate.net/publication/227441142_Logistic_regression_in_data_analysis_An_overview

(PDF) Multilingual Cyberbullying Detection System. (n.d.). Retrieved April 14, 2024, from https://www.researchgate.net/publication/335793614_Multilingual_Cyberbullying_Detection_System

(PDF) Performance Analysis of Boosting Classifiers in Recognizing Activities of Daily Living. (n.d.). Retrieved April 14, 2024, from https://www.researchgate.net/publication/339157430_Performance_Analysis_of_Boosting_Classifiers_in_Recognizing_Activities_of_Daily_Living

(PDF) Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language. (n.d.). Retrieved April 14, 2024, from https://www.researchgate.net/publication/339174318_Presumptive_Detection_of_Cyberbullying_on_Twitter_through_Natural_Language_Processing_and_Machine_Learning_in_the_Spanish_Language

(PDF) Scalable and timely detection of cyberbullying in online social networks. (n.d.). Retrieved April 14, 2024, from https://www.researchgate.net/publication/326166411_Scalable_and_timely_detection_of_cyberbullying_in_online_social_networks

(PDF) Text Classification Algorithms: A Survey. (n.d.). Retrieved April 16, 2024, from https://www.researchgate.net/publication/332463886_Text_Classification_Algorithms_A_Survey

(PDF) XBully: Cyberbullying Detection within a Multi-Modal Context. (n.d.). Retrieved April 14, 2024, from https://www.researchgate.net/publication/329302224_XBully_Cyberbullying_Detection_within_a_Multi-Modal_Context

Pennington, J., Socher, R., & Manning, C. (2014). *Glove: Global Vectors for Word Representation*. *14*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Ptaszynski, M., Dybala, P., Matsuba, T., Masui, F., Rzepka, R., & Araki, K. (2010). *Machine Learning and Affect Analysis Against Cyber-Bullying*.

Rafiq, R. I., Hosseinmardi, H., Han, R., Lv, Q., Mishra, S., & Arredondo Mattson, S. (2015). *Careful what you share in six seconds: Detecting cyberbullying instances in Vine*. 617–622. https://doi.org/10.1145/2808797.2809381

Raisi, E., & Huang, B. (2016). Cyberbullying Identification Using Participant-Vocabulary Consistency. *ArXiv*. https://www.semanticscholar.org/paper/Cyberbullying-Identification-Using-Consistency-Raisi-Huang/8813d9cf19e201bf81e1d919a098e7f5921954e3

Raza, M., Memon, M., Bhatti, S., & Bux, R. (2020). *Detecting Cyberbullying in Social Commentary Using Supervised Machine Learning* (pp. 621–630). https://doi.org/10.1007/978-3-030-39442-4_45

Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using Machine Learning to Detect Cyberbullying. *2011 10th International Conference on Machine Learning and Applications and Workshops*, 241–244. 2011 Tenth International

Conference on Machine Learning and Applications (ICMLA 2011). https://doi.org/10.1109/ICMLA.2011.152

Roy, P., Tripathy, A., & Das, T. (2022). Cyberbullying detection: An ensemble learning approach. *International Journal of Computational Science and Engineering*, *25*. https://doi.org/10.1504/IJCSE.2022.10047721

Rybnicek, M., Poisel, R., & Tjoa, S. (2013). *Facebook Watchdog: A Research Agenda For Detecting Online Grooming and Bullying Activities*. 2854–2859. https://doi.org/10.1109/SMC.2013.487

Sahlgren, M., Isbister, T., & Olsson, F. (2018). Learning Representations for Detecting Abusive Language. In D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, & J. Wernimont (Eds.), *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)* (pp. 115–123). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5115

Saini, H., Mehra, H., Rani, R., Jaiswal, G., Sharma, A., & Dev, A. (2023). Enhancing cyberbullying detection: A comparative study of ensemble CNN–SVM and BERT models. *Social Network Analysis and Mining*, *14*(1), 1. https://doi.org/10.1007/s13278-023-01158-w

Salin, D. (2003). Ways of Explaining Workplace Bullying: A Review of Enabling, Motivating and Precipitating Structures and Processes in the Work Environment. *Human Relations*, *56*(10), 1213–1232. https://doi.org/10.1177/00187267035610003

Salin, D. (2008). The prevention of workplace bullying as a question of human resource management: Measures adopted and underlying organizational factors.

Scandinavian Journal of Management, *24*(3), 221–231. https://doi.org/10.1016/j.scaman.2008.04.004

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-Centric Computing and Information Sciences*, *10*(1), 1. https://doi.org/10.1186/s13673-019-0205-6

Schonfeld, A., McNiel, D., & Toyoshima, T. (2023). Cyberbullying and Adolescent Suicide. *The Journal of the American Academy of Psychiatry and the Law*, *51*(1).

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry*, *49*(4), 376–385. https://doi.org/10.1111/j.1469-7610.2007.01846.x

Snakenborg, J., Van Acker, R., & Gable, R. (2011). Cyberbullying: Prevention and Intervention to Protect Our Children and Youth. *Preventing School Failure*, *55*, 88–95. https://doi.org/10.1080/1045988X.2011.539454

Sonnentag, S., & Fritz, C. (2015). Recovery from job stress: The stressor-detachment model as an integrative framework. *Journal of Organizational Behavior*, *36*(S1), S72–S103. https://doi.org/10.1002/job.1924

Stauffer, S., Heath, M., Coyne, S., & Ferrin, S. (2012). High school teachers' perceptions of cyber bullying: Prevention and intervention strategies. *Psychology in the Schools*, *49*, 353–367. https://doi.org/10.1002/pits

Tarwani, S., Jethanandani, M., & Kant, V. (2019). *Cyberbullying Detection in Hindi-English Code-Mixed Language Using Sentiment Classification* (pp. 543–551). https://doi.org/10.1007/978-981-13-9942-8_51

Teixeira, A., Ferreira, T., & Borges, E. (2016). Bullying no trabalho: Perceção e impacto na saúde mental e vida pessoal dos enfermeiros. *Revista Portuguesa de Enfermagem de Saúde Mental*. https://doi.org/10.19131/rpesm.0128

Tenenbaum, L., Varjas, K., Meyers, J., & Parris, L. (2011). Coping strategies and perceived effectiveness in fourth through eighth grade victims of bullying. *School Psychology International*, *32*, 263–287. https://doi.org/10.1177/0143034311402309

Types of Cyberbullying—Examples of Bullying Online. (n.d.). Social Media Victims Law Center. Retrieved July 26, 2024, from https://socialmediavictims.org/cyberbullying/types/

van der Zwaan, J., Dignum, V., & Jonker, C. (2012). A Conversation Model Enabling Intelligent Agents to Give Emotional Support. In *Studies in Computational Intelligence* (Vol. 431, pp. 47–52). https://doi.org/10.1007/978-3-642-30732-4_6

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., Pauw, G., Daelemans, W., & Hoste, V. (2015, September 7). *Detection and fine-grained classification of cyberbullying events*.

Waseem, Z., & Hovy, D. (2016). *Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter*. 88–93. https://doi.org/10.18653/v1/N16-2013

Webb, M., Burns, J., & Collin, P. (2008). Providing online support for young people with mental health difficulties: Challenges and opportunities explored. *Early*

Intervention in Psychiatry, *2*(2), 108–113. https://doi.org/10.1111/j.1751-7893.2008.00066.x

Yang, L.-Q., Caughlin, D. E., Gazica, M. W., Truxillo, D. M., & Spector, P. E. (2014). Workplace mistreatment climate and potential employee and organizational outcomes: A meta-analytic review from the target's perspective. *Journal of Occupational Health Psychology*, *19*(3), 315–335. https://doi.org/10.1037/a0036905

Ybarra, M., & Mitchell, K. (2007). Prevalence and Frequency of Internet Harassment Instigation: Implications for Adolescent Health. *The Journal of Adolescent Health : Official Publication of the Society for Adolescent Medicine*, *41*, 189–195. https://doi.org/10.1016/j.jadohealth.2007.03.005

Ybarra, M., Mitchell, K., Wolak, J., & Finkelhor, D. (2006). Examining Characteristics and Associated Distress Related to Internet Harassment: Findings From the Second Youth Internet Safety Survey. *Pediatrics*, *118*, e1169-77. https://doi.org/10.1542/peds.2006-0815

Yin, D., Xue, Z., Hong, L., Davison, B., Edwards, A., & Edwards, L. (2009). *Detection of harassment on Web 2.0*.

Zapf, D., & Einarsen, S. (2005). *Mobbing at Work: Escalated Conflicts in Organizations.* https://doi.org/10.1037/10893-010

Zapf, D., & Gross, C. (2001). Conflict escalation and coping with workplace bullying: A replication and extension. *European Journal of Work and Organizational Psychology*, *10*(4), 497–522. https://doi.org/10.1080/13594320143000834

Zhao, R., Zhou, A., & Mao, K. (2016, January 4). *Automatic Detection of Cyberbullying on Social Networks based on Bullying Features*. https://doi.org/10.1145/2833312.2849567

Zinovyeva, E., Härdle, W. K., & Lessmann, S. (2020). Antisocial online behavior detection using deep learning. *Decision Support Systems*, *138*, 113362. https://doi.org/10.1016/j.dss.2020.113362

Zohar, D., & Luria, G. (2005). A Multilevel Model of Safety Climate: Cross-Level Relationships Between Organization and Group-Level Climates. *Journal of Applied Psychology*, *90*(4), 616–628. https://doi.org/10.1037/0021-9010.90.4.616

13. Appendices

Appendics section has a separate file (Thesis_Appendics.pdf) and link to appendices:

https://drive.google.com/file/d/1YjF_3hzRX4ovH1x0tqdJhK1_hRrxoPVq/view?usp=sharing

14. Glossaries

- ADASYN: Adaptive Synthetic
- AdaBoost: Adaptive Boosting
- API: Application Programming Interface
- AUC: Area Under the Curve
- Bagging: Machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms
- BERT: Bidirectional Encoder Representations from Transformers
- Bi-LSTM: Bidirectional Long Short-Term Memory
- Boosting: Ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning
- CNN: Convolutional Neural Network
- Confusion Matrix: Table used to describe the performance of a classification model
- Corpus: Large and structured set of texts
- Cross-validation: Model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set
- CSS: Cascading Style Sheets
- Cyberbullying: Use of electronic communication to bully a person
- Decision Tree: Tree-like model of decisions and their possible consequences
- Discourse Analysis: Study of the use of language in communication
- DM: Direct Message
- Ensemble Learning: Use of multiple learning algorithms to obtain better predictive performance
- F1 Score: Harmonic mean of precision and recall
- Feature Extraction: Process of reducing the number of resources required to describe a large set of data accurately
- Feature Importance: Technique used in machine learning to determine which features in a dataset have the most impact on the prediction
- GAM: General Aggression Model
- GloVe: Global Vectors for Word Representation
- GOSS: Gradient-based One-side Sampling

- Gradient Descent: First-order iterative optimization algorithm for finding a local minimum of a differentiable function
- GridSearchCV: Grid Search Cross-Validation
- HTML: Hypertext Markup Language
- Hyperparameter: Parameter whose value is set before the learning process begins
- IM: Instant Message
- JS: JavaScript
- Kernel: Function used in support vector machines
- Lemmatization: Process of reducing words to their base or dictionary form
- Lexical Analysis: Process of converting a sequence of characters into a sequence of tokens
- LightGBM: Light Gradient Boosting Machine
- Logistic Regression: Statistical method for predicting a binary outcome
- LSTM: Long Short-Term Memory
- ML: Machine Learning
- Morphology: Study of words, how they are formed, and their relationship to other words in the same language
- Naive Bayes: Probabilistic classifier based on applying Bayes' theorem
- Neural Network: Computing system inspired by the biological neural networks that constitute animal brains
- NLP: Natural Language Processing
- NLTK: Natural Language Toolkit
- Overfitting: Modeling error that occurs when a function is too closely fit to a limited set of data points
- PCA: Principal Component Analysis
- Phonology: Branch of linguistics concerned with the systematic organization of sounds in spoken languages
- Pragmatic Analysis: Study of how context contributes to meaning
- Precision: Ratio of true positive predictions to the total predicted positives
- PSC: Psychosocial Safety Climate
- Psychosocial: Relating to the interrelation of social factors and individual thought and behavior

- Recall: Ratio of true positive predictions to the total actual positives
- RF: Random Forest
- ROC: Receiver Operating Characteristic
- Semantic Analysis: Process of relating syntactic structures to their language-independent meanings
- Semi-supervised Learning: Machine learning task that uses both labeled and unlabeled data for training
- Sentiment Analysis: Process of computationally identifying and categorizing opinions expressed in a piece of text
- SGD: Stochastic Gradient Descent
- SLT: Social Learning Theory
- SMOTE: Synthetic Minority Over-sampling Technique
- Stemming: Process of reducing words to their root form
- Supervised Learning: Machine learning task of learning a function that maps an input to an output based on example input-output pairs
- SVG: Scalable Vector Graphics
- SVM: Support Vector Machine
- Syntactic Analysis: Process of analyzing a string of symbols conforming to the rules of a formal grammar
- TextBlob: Python library for processing textual data
- TF-IDF: Term Frequency-Inverse Document Frequency
- Tokenization: Process of breaking down text into individual units
- t-SNE: t-Distributed Stochastic Neighbor Embedding
- UCI: University of California, Irvine
- Underfitting: Modeling error that occurs when a function is too simple to capture the underlying structure of the data
- Unsupervised Learning: Machine learning task of inferring a function that describes the structure of unlabeled data
- URL: Uniform Resource Locator
- Word2Vec: Word embedding technique
- XGBoost: Extreme Gradient Boosting