International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I3P105

Volume 74 Issue 3, 54-74, March 2026

Original Article

Severity Detection of Cyberbullying in Saudi-Dialect Tweets: A Machine-Learning Approach

Bader Azi Alanazi[1], Chin-Teng Lin[2]

[1]*University of Technology Sydney (Australia).*
Jouf University (Kingdom of Saudi Arabia)
[2]*University of Technology Sydney (Australia).*

[1]*Corresponding Author : baenzi@ju.edu.sa*

Received: 06 March 2025 Revised: 24 January 2026 Accepted: 29 January 2026 Published: 28 March 2026

Abstract *- Social media platforms such as Twitter (known as X) have become channels for global communication, but have also led to an increase in cyberbullying, which carries serious psychological risks. Although much existing research has focused on detecting cyberbullying in English, there is an apparent lack of studies addressing this issue in Arabic, particularly for severity classification. This study aims to evaluate machine learning classifiers trained on balanced, pre-processed Saudi dialect data for four-level cyberbullying severity detection (non-cyberbullying, low, medium, and high) and to assess the impact of systematic class balancing on minority class performance. The study applied Support Vector Machine (SVM) and Naïve Bayes (NB) classifiers, using Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction. A dataset of 5,819 Saudi-dialect tweets was annotated into four severity categories and evaluated across 28 experimental scenarios combining different pre-processing tools (CAMeL, NLTK, Araby) and balancing techniques (random insertion, random oversampling, synonym replacement). The highest accuracy of 92.23% was achieved using BoW+SVM with NLTK pre-processing and stop word removal, representing a 27.43% absolute improvement over the imbalanced baseline of 64.80% accuracy. Random oversampling proved to be the most effective, accounting for 96-99% of the performance gains. Per-class F1-scores ranged from 0.88 (low severity) to 0.95 (high severity and non-cyberbullying), providing further evidence of the importance of balanced training data for achieving reliable performance across all severity levels. To the best of the authors' knowledge, this is the first study to implement four-class cyberbullying severity detection for Saudi dialect tweets.*

Keywords *- Text Classification, Machine Learning, Cyberbullying Detection, Arabic social media, Saudi dialect, Support Vector Machine(SVM), Naïve Bayes (NB).*

1. Introduction

Social media platforms, such as Twitter and Facebook, have changed interpersonal communication and information exchange. Twitter alone records an estimated 330 million monthly active users worldwide [1]. According to [2], Twitter has considerable reach within the Kingdom of Saudi Arabia, attracting 14.4 million users in 2025, with a continuous increase expected to reach 15.08 million by 2028. The increased usage of the platform emphasises its importance in Saudi society as a communication and information-sharing medium. The population of Saudi Arabia, as reported by [3], is 27 million, and 20% is aged 10-19 years. Twitter and other social platforms are embedded in daily life, enabling rapid news sharing, information exchange, and public expression of opinions. Pseudonymity allows users to mask their identities, offering privacy but weakening accountability [4]. These affordances and large user bases have increased opportunities for misuse, including harassment, rumour spreading, and false postings. Cyberbullying is not confined to social platforms; it also occurs via emails and instant-messaging applications. It manifests as spreading rumours, leaking or disclosing private information, and publishing insulting or fabricated content [5]. These behaviours are associated with anxiety and depression and, in extreme cases, suicidal ideation among victims [6]. Freed from spatial constraints, online harassment can shadow individuals around the clock. The frequent use of social media platforms by teenagers makes them vulnerable to cyberbullying [4, 7]. Accordingly, focused measures are needed to prevent cyberbullying and protect users from risks. As social media participation grows, cyberbullying rises in tandem, creating a clear need for automated detection tools. Early efforts show promise: an Arabic SVM-based detector reported 95.74% accuracy, outperforming a Naïve Bayes baseline [8]; another study achieved 95.9% with Naïve Bayes on Twitter-YouTube data[9]; and a 2021 two-tier system that first filtered violent content again found SVM to be the strongest [10]. Nevertheless, most published systems target English, and research on Arabic remains comparatively sparse

[8]. Therefore, scalable Arabic-language detectors are needed to support timely responses from platforms, governments, and educators.

Arabic presents additional challenges for automation, such as rich morphology and extensive dialectal variation. Models trained for one dialect rarely apply well to others because vocabulary, spelling conventions, and idioms differ across regions. Cultural context further complicates classification, as terms acceptable in one community may be offensive in another [9]; for instance, words such as "كلب" (dog) or "حمار" (donkey) may be acceptable in some contexts but not others [8]. Due to these, linguistic and cultural factors make dialect-specific corpora and evaluation benchmarks essential. Broadening such resources will improve detection accuracy and help prevent harmful content across Arabic social media.

Automatic cyberbullying detection, particularly in English, has been a growing interest recently. However, an earlier literature review on automated cyberbullying detection identified several limitations in previous studies. First, research on Arabic content is lacking. Existing research lacks assessments of cyberbullying intensity specifically in Saudi dialects. Third, several of these studies used an imbalanced dataset, which may have led to biased results. These identified gaps underscore the absence of an automated system specifically designed to detect cyberbullying severity in the Saudi dialect, trained on appropriately balanced data. This study addresses these limitations by examining whether balanced and pre-processed data are critical for effective multi-class severity detection in Saudi dialects of cyberbullying. Specifically, it investigates whether pre-processing and class balancing enable machine-learning classifiers to accurately distinguish between the four severity levels.

This paper presents a system for detecting the severity of cyberbullying in tweets written in the Saudi dialect. In the approach, Support Vector Machine (SVM) and Naïve Bayes (NB) classifiers are used with Bag-of-Words and TF-IDF features to capture signals relevant to severity in Saudi Arabic.

The Contributions of this study are as follows: (1) categorises cyberbullying severity in Saudi dialect tweets into four classes (High, Medium, Low, Non-cyberbullying); (2) constructs and evaluates 28 scenarios combining pre-processing pipelines (CAMeL, NLTK, Araby, and stop word removal) with class-balancing techniques (random insertion, random oversampling, and synonym replacement); (3) shows that balancing the dataset strengthens the recognition of severity, with BoW + linear SVM achieving the best performance under stop word removal; and (4) provides a clear, reproducible evaluation protocol (data splits, metrics, and implementation details). Collectively, prior work in Arabic often treats cyberbullying as binary detection or general offensive language classification. The present study targets severity-aware classification using four labels in the Saudi dialect. It examines 28 scenario configurations to identify pre-processing and balancing choices that significantly affect severity. To the best of the authors' knowledge, this is the first study to implement a four-class cyberbullying severity detection model for Saudi dialect tweets.

The remainder of this paper is structured as follows: Section 2 provides the background. Section 3 reviews related work with an emphasis on Arabic cyberbullying detection. Section 4 details the methodology, including data collection, pre-processing, oversampling, feature extraction, and classifiers. Section 5 describes the experimental setup and reports the results. Section 6 concludes with future work.

2. Background

2.1. Cyberbullying

2.1.1. The Definition of Cyberbullying

The accurate definitions of cyberbullying are essential for designing effective solutions. According to [11], cyberbullying is defined as online harassment through social media, email, or instant messaging. A cyberbully is an individual or group that intentionally commits aggression against a target that cannot defend itself. Similarly, researchers [12-15] have defined cyberbullying as the purposeful use of social media, such as Twitter, to harm a person or group. The cyberbullying behaviours involved, outlined further in this work, take several harmful forms. Importantly, online harassment can persist without pause-twenty-four hours a day, seven days a week-whereas traditional bullying typically ends once the in-person interaction ends.

2.1.2. Impact of Cyberbullying

Several researchers have described cyberbullying as an epidemic. Brunecz [16] notes that it "has become an epidemic in today's society". Paolini [17] similarly warns of a "silent-killer epidemic" affecting children and adolescents, and Zhange [18] refers to "the silent epidemic of cyberbullying". Cyberbullying threatens psychological health, self-esteem, and social relationships and can cause isolation and long-term emotional suffering [19]. Studies have found that victims of cyberbullying frequently experience anxiety, depression, and suicidal ideation [20]. The consequences impact school grades, friendships, home life, school, and community. One survey revealed that 31% of school students experienced cyberbullying at least once [21]. In the United States, impacted students had higher absences and switched to homeschooling [22, 23]. A survey in the United Kingdom [24] that involved 2,218 secondary students aged 11-19 found that one in four (25%) had experienced cyberbullying. Further research conducted in Saudi Arabia confirms that cyberbullying is a serious issue. In one example [25], a survey of 761 high school students (ages 15-19) from Riyadh found that 18% of the students were victims of cyberbullying. The

authors defined cyberbullying as an issue of rising concern that is harmful to mental health and academic achievement. In another study [26] from Jazan, a survey conducted among 355 students (12-18 years old) found a high prevalence of cyberbullying, at 42.8%. The authors recommended collaboration to protect adolescents. At King Saud University, a new study [27] of 203 female undergraduates from 12 colleges found that 41.6% of the students experienced cyberbullying. The National Family Safety Programme in Saudi Arabia launched a 2014 campaign to address cyberbullying and raise public awareness for students who are affected [8].

2.1.3. The Types of Cyberbullying

Cyberbullying on Twitter and other social media platforms involves malicious information gathering, harassment, and intimidation. Studies have identified numerous forms it can take [6, 28-33]. The forms of cyberbullying aim to harass, threaten, and target victims on online platforms. These forms include:

- Trickery: A user tricks the victim into sharing personal or private information about themselves, which is later used as a form of online harassment [6, 32].
- Harassment: The user sends numerous messages to another person containing offensive content [29].
- Exclusion: This kind of bullying is used by teens and young adults, who stop or do not let a particular person join their group online [32].
- Flaming: Online arguments that involve people insulting or flaming each other, using abusive language and comments as well [31, 32].
- Outing: This form of bullying involves publishing private or embarrassing information, photos, or videos online [33].
- Flooding: The user sends harassing messages/comments to the victim repeatedly [28].
- Cyberstalking: Publishing information about the victim online to spread rumours, lies, and electronic threats against the person on the Internet [32].
- Denigration: False and damaging statements made about an individual's character or reputation [32, 33].
- Masquerading: Pretending to be someone else to spread rumours, lies, and information against the victim to harm their reputation [32].
- Trolling: Provoking others for fun, causing arguments, and starting conflicts [30, 33].

2.2. Machine Learning Techniques

The increasing popularity of the Internet and the use of social media platforms have required the development of mechanisms to detect cyberbullying. Machine learning has been helpful in this regard, where supervised learning using algorithms such as Naïve Bayes or Support Vector Machines helps classify data and identify harmful content based on labelled training data [34, 35]. Unsupervised learning techniques such as K-means clustering have also been used to detect implied patterns of bullying [36]. Semi-supervised learning approaches have been employed to effectively combine supervised and unsupervised learning to detect bullying with minimal training data [37, 38]. Furthermore, deep learning techniques such as Long Short-Term Memory Networks (LSTMs) and Convolutional Neural Networks (CNNs) are used to detect cyberbullying by analysing large amounts of social media data to uncover deeper patterns and contexts [7, 39-43].

2.3. The Arabic Language and Natural Language Processing

Natural language processing (NLP) is a type of artificial intelligence that focuses on interpreting human language [44, 45]. NLP focuses on computational approaches to process human language through its written and spoken forms and aims to enable computers to understand it. It also involves studying syntax, grammar, and semantics to interpret language, which are required for meaning representation and extraction [46, 47]. NLP is used for various applications such as automation in translation, speech recognition, and sentiment analysis by automatically converting unstructured text to structured data [44, 46, 48]. Arabic is one of the six official languages of the United Nations, and millions of people speak Arabic. It has a complex morphology, multiple dialects, and is written from right to left [49, 50]. It also includes 28 letters in different styles [51, 52]. Arabic has two forms, Standard and Dialectal. The standard is divided into Classical Arabic (CA) and Modern Standard Arabic (MSA). Dialectal Arabic is represented in all spoken languages across Arabic countries [50, 53]. In recent years, significant progress has been made in Arabic NLP, including the development of tools and resources for various tasks, such as sentiment analysis and text categorisation, as well as improvements in Arabic text processing [8, 54, 55].

3. Related Work

This part provides an overview of the previous study of the cyberbullying system, including summaries of the findings, the performance of each investigation, and the most accurate results. In 2017 [56], they proposed a machine learning approach using the Support Vector Machine (SVM) and Naïve Bayes (NB) algorithms to detect cyberbullying on Arabic social media. They collected data from Facebook and Twitter to conduct the research. The experimental results showed that the machine learning algorithm was able to detect Arabic cyberbullying with SVM achieving 0.934 and NB achieving 0.901. Another study [57] implemented predictive modelling to detect antisocial behaviour in Arabic YouTube comments. By employing a large dataset of offensive and non-offensive Arabic comments, researchers trained their model using a Support Vector Machine (SVM) classifier. They achieved an impressive accuracy rate of 90.05% in identifying such behaviours.

In 2019, researchers [9] introduced an automated machine learning-based method for detecting cyberbullying in Arabic. They applied the Naïve Bayes classifier algorithm to train their model using authentic data from major social media platforms, such as Twitter and YouTube. The results were promising, with an accuracy rate of 95.9%, further supporting the effectiveness of machine learning in identifying Arabic cyberbullying. In 2021, [10] introduced a supervised machine learning method to build a two-tier classification system for violent Arabic texts. The initial tier differentiated between violent and non-violent content, whereas the second tier categorised violent text as either cyberbullying or threatening. The authors used the SVM and NB algorithms to test feature extraction methods and stop word removal settings. The results highlighted that SVM performed better than NB, establishing its effectiveness in this area.

In 2023, a study [8] presented a machine-learning approach for detecting cyberbullying on Arabic social media. The authors compared support vector machine (SVM) and naïve Bayes classifiers and highlighted the linguistic complexity of Arabic and varied user interactions as key challenges. SVM outperformed naïve Bayes, achieving an accuracy of 95.74 per cent.

In 2024 [58], researchers built a supervised pipeline for Arabic tweets that cleans and normalises the text, extracts TF-IDF or unigram features, and evaluates nine classifiers: SVM, naïve Bayes, random forest, logistic regression, Bagging, AdaBoost, gradient boosting, LightGBM, and XGBoost. XGBoost combined with TF-IDF recorded the best overall results, achieving an accuracy of 89.95% and an F1-score of 88.82%. Building on prior research, in 2017, researchers [56] applied SVM and NB classifiers to Facebook and Twitter data, achieving effectiveness scores of 0.934 and 0.901. Subsequent studies have refined these methods; a 2018 study [57] achieved 90.05% accuracy with SVM on YouTube comments. By 2019, an automated NB approach had reached 95.9% accuracy [9], and in 2021, a two-tier system using an SVM demonstrated superior performance [10]. In 2023, the use of SVM and NB classifiers led to a significant accuracy rate of 95.742% [8]. Recently, researchers [58] in 2024 applied nine classifiers to Arabic tweets, and XGBoost outperformed the others with 89.95% accuracy.

The findings across the studies show that Support Vector Machines (SVM) mostly outperform other classifiers in detecting Arabic cyberbullying, indicating that SVM models capture the language's subtle abusive cues more effectively. However, automated Arabic cyberbullying research remains in its early stages, and previous studies have not evaluated the severity of cyberbullying in the Saudi dialect. This study aims to fill this gap by developing a system capable of detecting the severity of cyberbullying in the Saudi dialect.

Table 1 contrasts representative Arabic cyberbullying / offensive language studies with the present work regarding dataset source pre-processing techniques, representations, classifiers, metrics, tasks, languages/dialects, and labels. The key difference here is the focus on multi-level severity in the Saudi dialect, explicit class balancing, and 28 scenario-based evaluations.

Table 1. Summary of related work

Ref.	Data source	Preprocessing Techniques	Representation	Classifier	Metrics	Task	Language/ Dialect	Label
[56]	Twitter (Arabic language 35,273), (English 91,431)	Data Cleaning, manual labelling, normalizing	TF-IDF, N-gram (WEKA pipeline)	SVM & NB	Precision, Recall and F-Measure	Cyberbullying detection	Arabic and English	Binary (bullying, non-bullying)
[57]	YouTube (15,050 comments)	Data Cleaning, Normalization, Tokenization	Bag-of-words, N-gram	SVM	Recall, Precision and F1-Score	Offensive language detection	Arabic	Binary (offensive vs non-offensive)
[9]	Twitter and YouTube (25,000 comments)	Data Cleaning, Normalization, Stemming	Not specified	NB	Precision, Recall, F-Measure and Accuracy	Cyberbullying detection	Arabic	Binary (bullying vs non-bullying)

[10]	Twitter (3,700 tweets)	Normalization, Noise Removal, Tokenization, Stop-word Removal	TF; embeddings (AraVec)	SVM & NB	Precision, Recall, F-Measure and Accuracy	Cyberbullying and threatening detection	Arabic (Saudi tweets)	Binary per-level: First (violent/ non-violent), Second (Cyberbullying/ threatening)
[8]	Twitter and YouTube (30,000 comments)	Data Cleaning, Normalization Stemmed, Segmented	TF-IDF, BoW	SVM & NB	Precision, Recall, F1-score and Accuracy	Cyberbullying detection	Arabic	Binary (bullying, non-bullying)
[58]	Twitter (9,000 tweets)	Normalization, Noise Removal, Tokenization, Removing Stop word	TF-IDF/Unigram	SVM, NB, RF, LR, Bagging, Boodt, Light, AdaBoost, and XGBoost	Precision, Recall, F1-score and Accuracy	Cyberbullying detection	Middle eastern regions with various dialects	Binary (bullying, non-bullying)

3.1. Multi-Class Severity Detection

Multi-class cyberbullying severity has been examined in several non-Arabic settings Table 2. Talpur and O'Sullivan [59] categorised four severity levels in English tweets using feature-based classical machine learning classifiers and reported strong overall performance (Random Forest accuracy of 93%).

Rahman et al. [60] introduce one of the first Bangla severity datasets and use XGBoost with PMI-SO and domain-specific features, reaching 87% accuracy, although under a highly skewed label distribution where non-cyberbullying is dominant. Wu and Tang [61] propose the HSAN model for three-level severity in Chinese social-media dialogues, but it still achieves only about 62.5% overall F1 score, with lower recall for the most serious class due to label imbalance. Meanwhile, Vyawahare and Govilkar [62] used XGBoost on a highly imbalanced English toxicity dataset with six severity-related labels, achieving an accuracy of 62.54%. In this study, the SVM configuration achieved 92.23% accuracy for four-class severity detection in Saudi dialect tweets, which falls within the upper range of these results, despite working with an under-resourced, morphologically rich dialect that features non-standard orthography and fewer NLP resources than those available for English.

Table 2. Multi-class detection of cyberbullying severity across different languages

Study	Language	severity	Model	Best score
[60]	Bangla (YouTube comments)	classes: no-bullying, low, medium, high	NB, SVM, LR, XGBoost	XGBoost accuracy 87%
[61]	Chinese (social-media dialogues)	3 classes: slight, medium, serious	HSAN (Hierarchical Squashing-Attention Network)	Overall F1 = 62.5%
[59]	English (Twitter)	4 classes: no-cyberbullying, low, medium, high	NB, SVM with RBF kernel, DT, RF, and KNN	RF Accuracy 93%**,**
[62]	English (Kaggle toxicity)	6 labels: toxic, severe toxic, obscene, threat, insult, identity hate	XGBoost, SVM, RF, DT	XGBoost Accuracy 62.54%
This study	Saudi dialect	4 classes: no-bullying, low, medium, high	SVM, NB	SVM 92.23% accuracy

Recent reviews have identified several limitations in cyberbullying detection systems. A literature review revealed that most of these systems use binary labels to detect cyberbullying and rarely address issues such as dialectical variation, class imbalance, or incidents. Azumah et al. [63] discuss challenges in dataset construction and evaluation, while Mahmud et al. [64] note the prevalence of English-targeted research and the lack of resources for low-resource languages. Allwaibed et al. [65] further emphasise the absence of dialect-specific resources in Arabic systems. The present

study addresses these issues by analysing Saudi dialect tweets at four severity levels instead of using binary classifications and systematically assessing 28 machine learning scenarios with different pre-processing and balancing strategies.

4. The Methodology

A machine-learning-based Saudi dialect tweets cyberbullying detection system was designed using the supervised classifiers Support Vector Machine (SVM) and Naïve Bayes (NB), as illustrated in Figure 1. The approach included: (1) data collection and cleaning, (2) pre-processing, (3) oversampling the minority class, (4) extracting features, (5) classification, and (6) evaluation. A method was developed to compare and quantify the importance of data balancing and preprocessing for severity classification. For that reason, 28 experiments were designed to evaluate all combinations of balancing techniques and preprocessing methods.

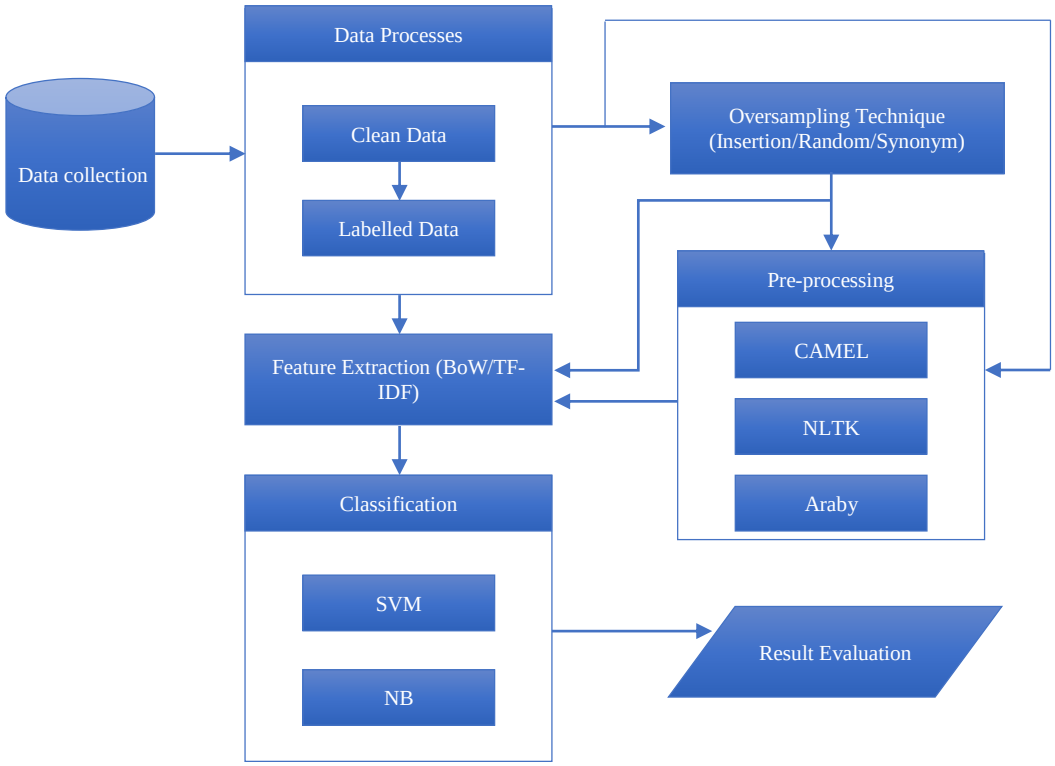

Fig. 1 Cyberbullying detection in the saudi dialect using ML

4.1. Data Collection and Processing

4.1.1. Data Collection

The data was collected using the official X (Twitter) Developer API. A Python script searched X using Saudi-focused hashtags. Search query options included filter and keyword options focused on cyberbullying and threat-based keywords. The data collection process yielded 11,384 tweets altogether. The datasets underwent de-duplication before being exported into Excel workbooks, where they received further processing and manual labelling for analysis.

4.1.2. Data Cleaning

The collected tweets were saved to an Excel file. A built-in Python script and manual filtering were used to remove noise and unnecessary terms. The first step was to build a Python script to filter out URLs, hashtags, "@mentions", numbers, and non-Arabic terms. Then, manual screening was performed to eliminate tweets that were not written in the Saudi dialect, contained advertisements, or included non-text material. The corpus size reached 5819 tweets suitable for classification once all filtering steps were completed.

4.1.3. Data Annotation

Based on the classification method in [66], the cyberbullying was categorised into four categories. Following this method, the tweets in the corpus were manually annotated into four classes: high, medium, low, and non-cyberbullying classes. The full dataset was labelled in three separate passes to strengthen the consistency of the annotations. Two native speakers of the Saudi dialect reviewed the annotation. A domain lexicon containing 3,340 cyberbullying-related words and phrases was compiled.

Tweets that targeted an individual's appearance or physical, or contained explicit sexual references, were labelled "high", reflecting their strong potential to cause psychological harm. Tweets that attacked a person or group based on race or religion were classified as "medium". Although still harmful,

tweets of minor or general insults were classified as "low". Finally, tweets free of any cyberbullying content were marked "non-cyberbullying," as shown in Table 3.

Table 3. Cyberbullying severity levels by category

Category	Severity level
Sexual/ Appearance	High
Political / Racial	Medium
General	Low
Non	Non-cyberbullying

To evaluate the consistency of the severity labels, the inter-annotator agreement was calculated across the complete set of 5,810 tweets. Pairwise Cohen's kappa between the three annotators ranged from $\kappa = 0.73$ to $\kappa = 1.00$, with an average $\kappa = 0.82$, signalling near-perfect agreement. The labels from the final pass were reviewed and approved by two native speakers of the Saudi dialect, and this validated set was used as the benchmark dataset for all experiments. The final class distribution is presented in Table 4.

Table 4. Distribution of annotated tweets by severity category

Category	Annotated Tweets
High	1587
Medium	221
Low	1299
Non-cyberbullying	2703

Table 5. Examples of tweets and severity classification.

Tweet	Translate	severity
أي شخص يحترمها أو يدافع عنها - سأبصق في وجهه.	Anyone who respects or defends her-I will spit in his face.	Low
مجموعة من الأوغاد يبالغون في تقدير بعضهم البعض - رغم أنوفكم.	A bunch of lowlifes hyping each other up-despite your snouts.	Medium
إنه يطاردها بلا مبالاة - ماذا تتوقع منه؟ *اسم*، ابن المرأة المنبوذة.	He is blindly chasing after her—what do you expect from him? *Name*, the outcast woman's son.	High
لأول مرة على تويتر، سورة البقرة كاملة في مقطع واحد؛ القارئ هو سعود الشريم. ضع التغريدة في المفضلة.	For the first time on Twitter, Surah Al-Baqarah is complete in one clip; the receiver is Saud Al-Shuraim. Put the tweet in your favourites.	Non-cyberbullying

4.2. Oversampling Technique for Balanced Data

Three oversampling techniques- random insertion, random oversampling, and synonym replacement were applied to balance the dataset. Balancing is critical because a well-distributed dataset allows the model to learn and predict each category accurately and without bias. These methods were selected for their effectiveness in textual data. The original class distribution was imbalanced, as summarised in Table 4.

First, Random insertion selects a synonym for a non-stop word in a sentence and inserts it at a random position; the procedure is repeated multiple times to bolster minority classes [67]. Random oversampling duplicates tweets from under-represented categories until each class contains an equal number of instances, preventing the model from favouring the majority class [68]. Third, Synonym replacement is a technique that randomly replaces some non-stop words with a randomly selected synonym, further augmenting minority classes [67].

Table 6 shows that the class distribution became balanced after oversampling. The augmentations were verified by randomly sampling 100 generated instances of each technique and manually checking them for any incorrect outputs. This step was to verify the dataset's accuracy, and the audited examples confirmed that the procedures generated accurate and appropriate instances.

Table 6. Distribution after oversampling

Category	Number
High	2703
Medium	2703
Low	2703
Non-cyberbullying	2703

A balanced dataset is important for model development because it promotes fair representation across classes and enables the model to learn effectively from all categories.

An even class distribution improves the accuracy and reliability of cyberbullying severity assessment by giving each severity level comparable influence during training and evaluation. This balanced foundation also increases confidence in subsequent analyses and strengthens the practical value of any interventions informed by the findings.

4.3. Data Pre-Processing

The Arabic language is morphologically rich, with complex rules and a large number of forms per word. The CAMeL [69], NLTK [70], and Araby [71] techniques were applied to Arabic words to assess their performance and select the best-performing technique for the language. The preprocessing techniques listed below were applied to all three techniques:

The first step, tokenisation, was used to break text into individual tokens; for example, the sentence "اليوم السهرة ياسلام" is tokenised as "السهرة, اليوم, ياسلام". Arabic Diacritics (also

called tashkeel or harakat) are the marks written above or below the letters of an Arabic word to signify the vowels or other pronunciation attributes such as doubled consonants. Diacritics were removed to reduce noise during preprocessing. An Arabic sentence with diacritics "السَّلَامُ عَلَيْكُم ورَحْمَةُ اللهِ وَبَرَكاتُهُ " would be " السلام عليكم ورحمة الله وبركاته " after stripping it of diacritics. Additionally, Arabic text was normalised, which involved converting various word forms into a standard form. For example, different forms of 'Alif' (أ, إ, آ) were replaced with a simple "ا". Thus, the word "أهلا" was normalised to "اهلا".

Additionally, repeated letters were removed to reduce words with repeated letters to their standard forms. For example, "رااااائع" was normalised to "رائع". After these steps, stop word removal was applied because such words occur frequently with no meaning or essential effect in sentences. In Arabic, examples include "و", "في", and "من". For example, the sentence with the stop word "الولد والبنت في المنزل" ("The boy and the girl are in the house"), once the stop words were removed, became "الولد البنت المنزل" ("The boy, the girl, the house"). Following the pre-processing steps, 28 scenarios were created from the dataset (Figure 2).

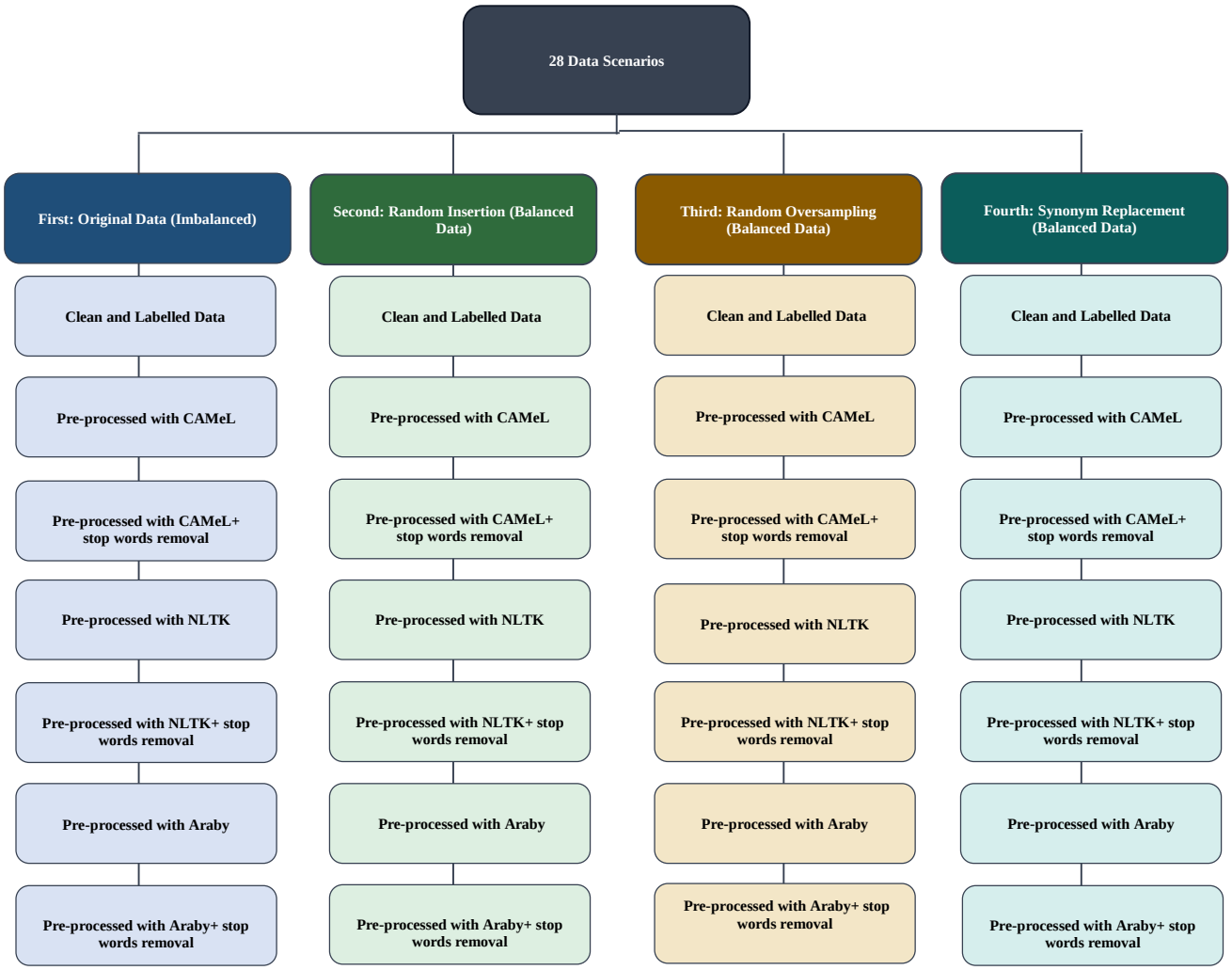

Fig. 2 Data scenarios

The dataset was divided into 4 main groups: one imbalanced group and three balanced groups using random insertion, random oversampling, and synonym replacement, respectively. For each of these groups, seven different scenarios were considered: (1) no pre-processing, (2) CAMeL, (3) CAMeL with stop word removal, (4) NLTK, (5) NLTK with stop word removal, (6) Araby, and (7) Araby with stop word removal. Combining the preprocessing approaches and balancing methods yielded 28 experimental scenarios, as shown in Figure 2. This experimental setting created multiple variants of the corpus. It allows for investigating the impact of pre-processing and data balancing in the latter stage of the study on the classification performance.

4.4. Feature Extraction

Two feature extraction algorithms were used to transform the data into machine learning readable formats. Term Frequency-Inverse Document Frequency and Bag-of-Words

are statistical measures used to extract the most informative tokens out of the total corpus of words in the document. Both are feature-extraction techniques that transform the input into a vector of numbers for learning algorithms [8].

TF-IDF is a measure of the weight (significance) of each token (term or word) that shows in a specific document of that specific corpus (collection of documents). The "TF" is obtained by determining the number of terms found in a particular document "d" and the number of times each term appears in a document 't', which is measured as follows:

$$TF_{t,d} = \frac{n_{t,d}}{\sum_k n_{t,d}} \quad (1)$$

Equation 1, where n_{t_d} is the frequency of the term "t" in a particular document 'd', and $\sum_k n_{t,d}$ is the number of times that terms occur in a document 'd'.

Inverse Document Frequency is an algorithm used to extract the most important keywords with the highest score calculated from two variables, TF and IDF, as follows:

$$IDF(t) = \log\left(\frac{1+N}{1+d_f(t)}\right) + 1 \quad (2)$$

where (N) is the total number of documents and $d_f(t)$ is the number of documents with the term "t" in it.

The TF is obtained as the multiplication of TF and IDF of a term, which is the most significant to a document of the corpus when compared to the corpus of documents:

$$TF - IDF_{t,d} = TF_{t,d} \times IDF(t) \quad (3)$$

Bag-of-Words (BoW) represents text with the word frequency of each word in a text in a vector of fixed size. The Tweets are considered to be input data, and every phrase's occurrence was computed. It is also stated that if the number of events is large, then calculating the word count for each word will yield the numerical value of that term as a vector.

In conclusion, feature extraction techniques converted the textual input data into a numerical form suitable for machine learning classifiers. The Term Frequency-Inverse Document Frequency (TF-IDF) algorithm assigns weights to the corpus based on the distribution of terms. Bag-of-Words (BoW) provides a basic word count vector of token occurrences. These approaches abstracted the most informative characteristics to enable fast, accurate learning of the dataset.

4.5. Classification

The study's baseline model consists of a linear-kernel Support Vector Machine (SVM) and Multinomial Naïve Bayes (NB) classifier. Linear SVM learns a maximal-margin hyperplane classifier in the input space, where text is represented as high-dimensional vectors. This classifier scales well when the number of features is large relative to the dataset size [34, 72]. The linear kernel allows the model to establish a linear decision boundary between classes, making it particularly suitable for high-dimensional data. SVMs are effective when the number of features exceeds the number of samples.

The multinomial NB applies Bayes' theorem under a conditional independence assumption, offering a probabilistic model that remains robust on sparse, high-dimensional data, even with limited training samples [34]. Analysing both models on the same corpus enables a direct evaluation of discriminative (hyperplane-based) versus generative (probabilistic) strategies for cyberbullying severity classification.

In this step, each of the 28 scenarios in Figure 2 was split into training and test sets using an 80:20 stratified split (train_test_split, random_state=42) to preserve the four severity classes (high, medium, low, and non-cyberbullying). Text was represented using unigram-bigram Bag-of-Words (CountVectorizer, ngram_range=(1,2)) and TF-IDF (TfidfVectorizer, same n-grams). Vectorisers were fitted only on the training split and then applied to the held-out test split to prevent data leakage. Two classifiers were considered: a linear Support Vector Machine (sklearn.svm.SVC(kernel='linear', probability=True)) with regularisation parameter C∈ {0.1,1,10}, tuned via three-fold cross-validation (best macro-F1 at C=1), and a Multinomial Naïve Bayes (MNB) classifier with Laplace smoothing α∈{0.01,0.1,1}. Severity labels were integer-encoded (non-cyberbullying=0, low=1, medium=2, high=3). Models trained on the 80% split were evaluated on the 20% split using sklearn.metrics with weighted averaging for accuracy, precision, recall, and F1-score. The experiments were performed using Python 3.10 and scikit-learn 1.5.0 on an Intel i7-11700 CPU (32 GB RAM). The main objective of this study was to examine how effectively the severity of cyberbullying can be detected in Saudi dialect tweets using well-established machine learning classifiers across a range of data balancing and pre-processing conditions. Therefore, deep learning architectures are considered a direction for future work, in which models such as CNNs, LSTMs, and BiLSTMs will be evaluated on the same corpus and scenarios to extend and compare them against the traditional baselines established in this study.

4.6. Evaluation Method

The model performance was assessed using four metrics: accuracy, F1-score, recall, and precision [73]. These metrics help provide a comprehensive assessment of the model's performance:

Accuracy: This metric measures the overall accuracy of the model.

$$Accuracy = \frac{Number\ of\ Correct\ Prediction}{Total\ Number\ of\ Predictions} \quad (4)$$

F1-score: This is the harmonic average of Precision and Recall, balancing the two.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision \times Recall} \quad (5)$$

Recall: measures the model's ability to identify all relevant instances.

$$Recall = \frac{True\ Positives}{True\ Positives \times False\ Negatives} \quad (6)$$

Precision: This metric measures the accuracy of the optimistic predictions made by the model.

$$Precision = \frac{True\ Positives}{True\ Positives \times False\ Positives} \quad (7)$$

These metrics provide an accurate and comprehensive evaluation of the model's ability to classify text data by severity of cyberbullying.

5. Results and Discussions

5.1. Results

This section presents the experimental results for detecting the severity of cyberbullying in Saudi dialect tweets. The original dataset, after the pre-processing and balancing pipelines, yielded 28 experiments, as shown in Figure 2. The following experiments evaluate the performance of the SVM and NB classifiers on the original imbalanced corpus and the three balanced corpora generated via random insertion, random oversampling, and synonym replacement.

Table 9 shows that, with systematic class balancing, accuracy increased from 64.80% for the imbalanced data to 92.23% with random oversampling. The balanced, pre-processed data enable more robust four-class severity detection. The results confirm the importance of pre-processing combined with class balancing to achieve reliable classification across all severity levels. The first dataset was imbalanced, consisting of 5810 tweets manually classified into four levels of severity (high, medium, low, and non-cyberbullying). Most of the tweets were labelled as non-cyberbullying and high severity Table 4.

The highest accuracy was 64.80% with BoW + SVM on the non-preprocessed corpus. The second-highest results, BoW + NB, TF-IDF + SVM, and TF-IDF + NB, were all in the 62.48% to 64.72% range. The CAMeL pipeline and stop-word removal yielded 67.99% using TF-IDF + SVM. In addition, using the pre-processing tools NLTK and Araby produced better performance with TF-IDF+NB and TF-IDF + SVM Table 7, achieving the highest accuracies of 67.47% and 66.61%, respectively.

Table 7. Best result in scenario 1 (original imbalanced data)

Model	Scenario	Accuracy
BoW+SVM	Original Data+No-preprocessing	64.80%
BoW+NB	Original Data+ Araby+ stop words removal	63.86%
TF-IDF +SVM	Original Data+ CAMeL+ stop words removal	67.99%
TF-IDF+NB	Original Data+ NLTK+ stop words removal	66.44%

The second group used the random insertion: the technique was applied to balance the data. After balancing, each category had 2703 tweets in total. The highest accuracy recorded was 90.11% for BoW + SVM, achieved when the CAMeL tool was combined with the stop-word removal technique Table 8. For all the other preprocessing techniques, improvements ranged from 82.52% to 90.06%.

Table 8. Best result in scenario 2 (random insertion balanced data)

Model	Scenario	Accuracy
BoW+SVM	Random Insertion balanced Data+CAMeL+ stop words removal	90.11%
BoW+NB	Random Insertion balanced Data+CAMeL+ stop words removal	84.28%
TF-IDF +SVM	Random Insertion balanced Data+NLTK+ stop words removal	88.49%
TF-IDF+NB	Random Insertion balanced Data+Araby+ stop words removal	83.73%

The third group used random oversampling to construct a balanced dataset containing 2,703 tweets per class. The highest performance was 92.23% with BoW + SVM and the NLTK pipeline, without stop word removal Table 9. The overall accuracy across all configurations in this group ranged from 84.74% to 92.23%. The results in this group indicate that a combination of balancing and pre-processing led to significant improvements over the imbalanced baseline.

Table 9. Best result in scenario 3 (random oversampling balanced data)

Model	Scenario	Accuracy
BoW+SVM	Random Oversampling balanced Data+NLTK+ stop words removal	92.23%
BoW+NB	Random Oversampling balanced Data+CAMeL	86.41%
TF-IDF +SVM	Random Oversampling balanced Data+NLTK+ stop words removal	90.43%
TF-IDF+NB	Random Oversampling balanced Data+CAMeL+ stop words removal	85.34%

The fourth group balanced data (Synonym Replacement): Synonym replacement produced a balanced dataset of 2,703 tweets per class. The highest accuracy of 90.43% was achieved with BoW+SVM under the NLTK pipeline without stop word removal Table 10. The accuracy of all configurations in this group ranged from 83.36% to 90.43%, indicating Performance gains from synonym-based balancing.

Table 10. Best result in scenario 4 (synonym replacement balanced data)

Model	Scenario	Accuracy
BoW+SVM	Synonym Replacement, balanced Data+NLTK+ stop words removal	90.43%
BoW+NB	Synonym Replacement, balanced Data+Araby+ stop words removal	84.84%
TF-IDF +SVM	Synonym Replacement, balanced Data+NLTK+ stop words removal	88.86%
TF-IDF+NB	Synonym Replacement balanced Data+CAMeL	84.56%

To examine the performance across individual severity levels, Table 11 presents the per-class precision, recall, and F1-scores for the best-performing configuration (BoW + linear SVM with random oversampling, NLTK preprocessing, and stop word removal). All four classes demonstrated high F1-scores, with values of approximately 0.90 for the non-cyberbullying and low-severity categories, and nearly perfect performance for the medium- and high-severity categories. The model consistently demonstrated a strong ability to detect all severity levels.

Table 11. Best scenario per-class metrics

	Precision	Recall	F1-score	Support
non-cyberbullying	0.855	0.943	0.897	564
low	0.886	0.882	0.884	501
medium	0.998	1	0.999	532
high	0.961	0.864	0.91	566

Figure 3 displays the confusion matrices. The majority of predictions are on the diagonal, as expected from the overall accuracy of 92.23%. Misclassifications occur between non-cyberbullying tweets predicted as low severity and some low/high severity tweets predicted as non-cyberbullying or misclassified as each other.

Medium-severity tweets were detected almost perfectly. In general, this suggests that the majority of misclassifications were due to borderline cases, where it was difficult to tell whether the tweet was low, high, or non-cyberbullying severity.

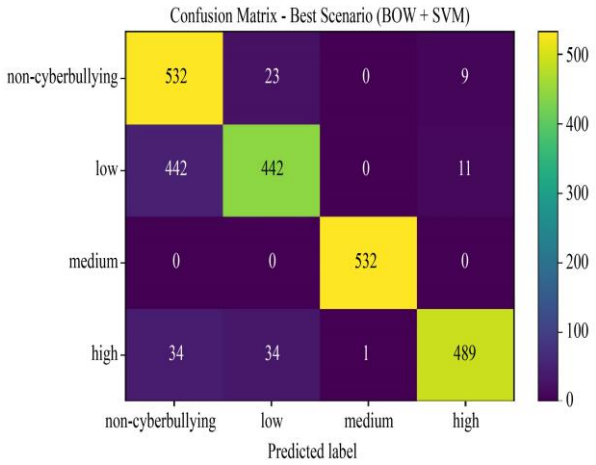

Fig. 3 Confusion matrix- best scenario

5.2. Discussion

The results of this study show that data balancing is essential for accurate multi-class severity detection. Data balancing yields both higher overall accuracy and improved results, as evidenced by a 27.43% performance improvement and uniform F1 scores ranging from 0.88 to 0.95. The evaluation of data-balancing techniques demonstrated improvements in model performance. The random oversampling with stop word removal was the most performant among the balanced data variants, achieving 92.23% accuracy. Analysis of balanced models indicated that training on balanced data yields better learning outcomes, reduces majority-class bias, and improves performance. The difference between the imbalanced and balanced cases also shows that the data balancing affected the model performance. The model performed poorly at classifying minority classes in imbalanced data, resulting in lower overall accuracy. On the other hand, balanced data significantly increased the model's accuracy in recognising patterns across different levels of severity. The findings of this study show that data-balancing techniques significantly improve the robustness of machine learning models. Improvement after pre-processing: Arabic text was pre-processed to address its rich, complex morphology, thereby improving classification accuracy using CAMeL, NLTK, and Araby. The Preprocessor tools consistently improved accuracy across all datasets. The highest accuracy of 92.23% for BoW + SVM was achieved through NLTK combined with stop word removal. The results of pre-processing are clearly evident after removing noise and irrelevant features, with a significant impact on the model's performance.

5.2.1. Model Performance and Interpretability

The BoW + SVM model showed superior performance compared to other models Table 12, especially after stop word removal. TF-IDF + SVM performed well with the CAMeL and NLTK tools, achieving a peak accuracy of 92.23% with BoW + SVM and stop-word removal. Comparing models' performance across scenarios revealed that BOW+SVM was

the most accurate, while TF-IDE+SVM excelled when combined with CAMeL and NLTK preprocessors. These results show that BOW and TF-IDF feature extraction techniques enhance model accuracy.

Table 12. Highest accuracy scenario per model

Model	Data scenario	Best Accuracy
BoW+SVM	Random Oversampling balanced Data+NLTK+ stop words removal	92.32%
BoW+NB	Random Oversampling balanced Data+CAMeL	86.41%
TF-IDF +SVM	Random Oversampling balanced Data+NLTK+ stop words removal	90.43%
TF-IDF+NB	Random Oversampling balanced Data+CAMeL+ stop words removal	85.34%

A linear SVM with BoW features is the best classifier; therefore, the decision boundary is easy to interpret: each word or n-gram receives a weight that indicates how much it adds to a particular severity class. An inspection of the most weighted features shows that direct personal attacks, insults, and sexual dominance have high severity predictions.

On the other hand, neutral or positive expressions are non-cyberbullying expressions. Cases of low and medium severity are linked to less harsh ridicule, criticism, and emotionally charged but less direct language use. These patterns are consistent with the severity annotation guidelines, suggesting that the classifier relies on linguistically meaningful information.

The results are organised in Table 13 to provide a clearer comparison of the model performance. The following structured presentation highlights the highest accuracy, F1-score, recall, and precision for easy comparison.

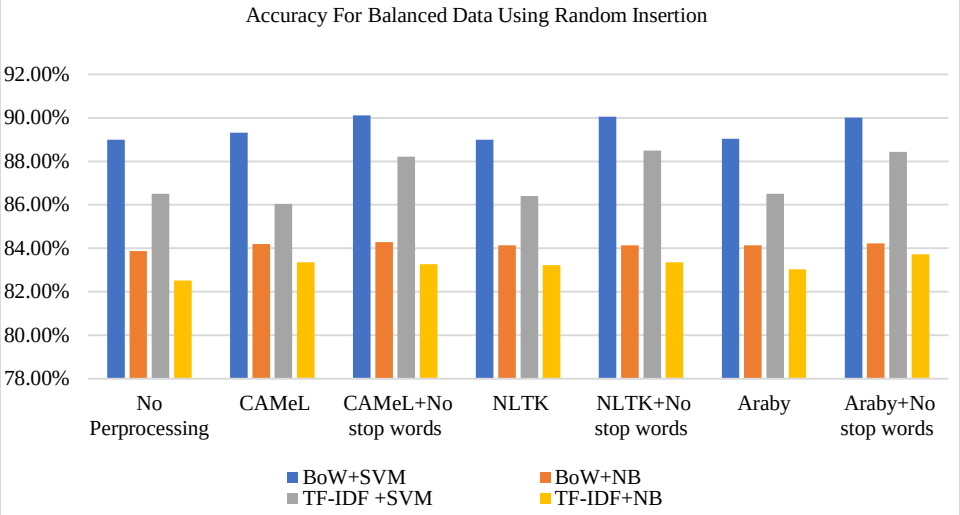

Fig. 4 Balanced data using random insertion

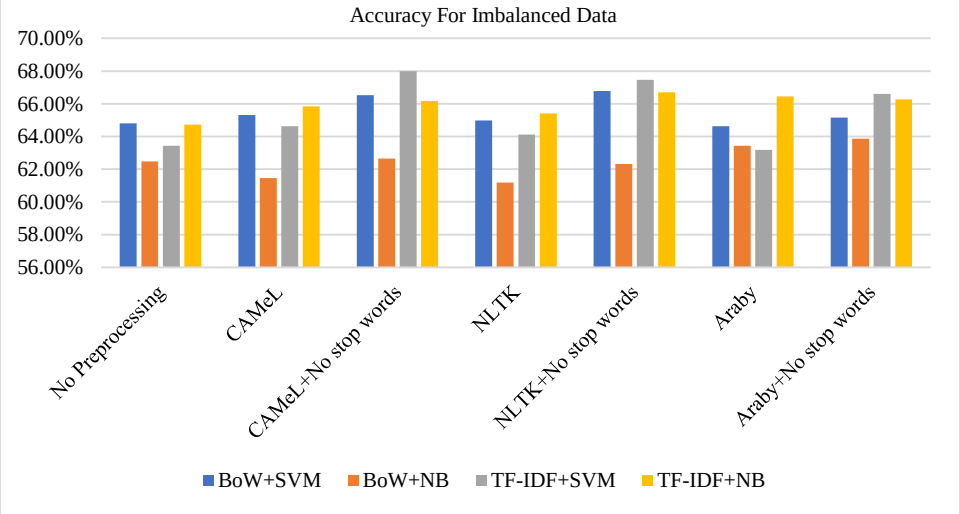

Fig. 5 Accuracy for imbalanced data

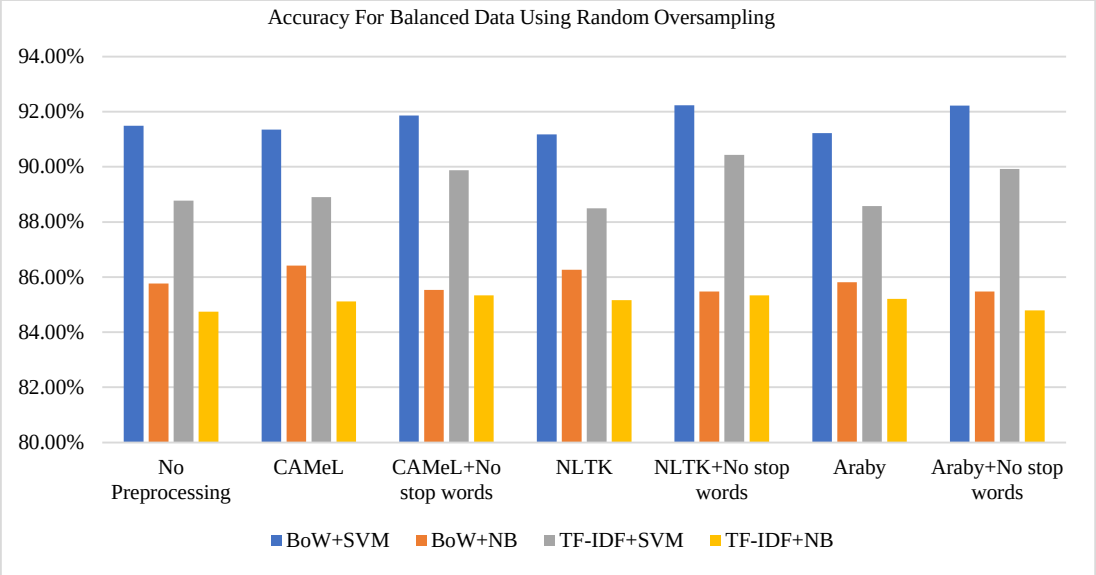

Fig. 6 Accuracy for balanced data using random oversampling

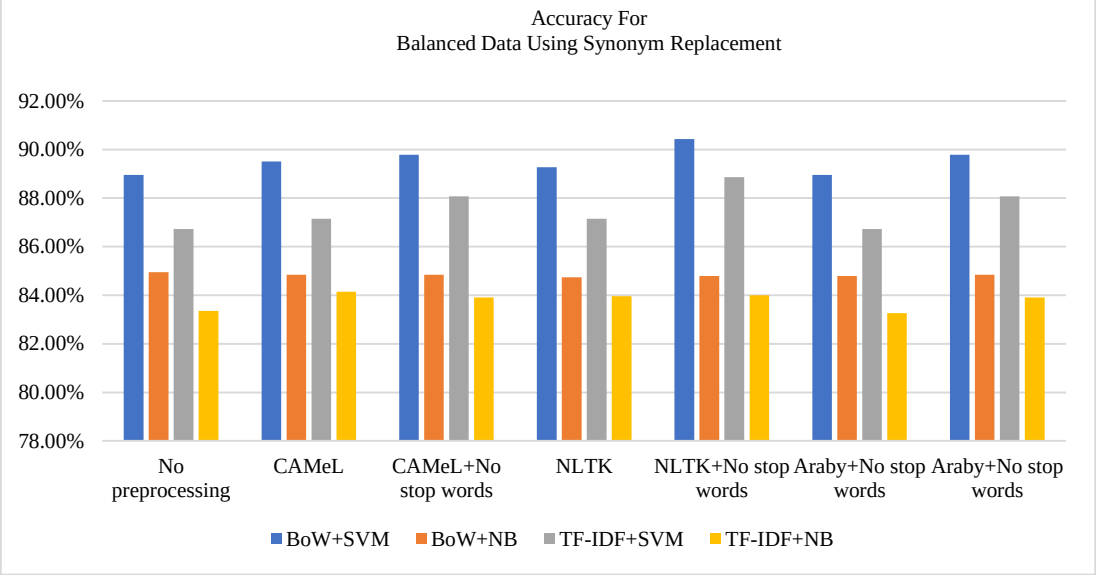

Fig. 7 Accuracy for balanced data using synonym replacement

The results of this study confirmed that data balancing and pre-processing are two critical steps for more accurate and objective models for detecting cyberbullying in the Saudi dialect. Figure 8 showed that these two steps have positively contributed to enhancing the model's performance by increasing the accuracy. Data balancing techniques have shown a positive impact on model performance. In this experiment, random oversampling yielded the highest accuracy, making it the most effective for handling class imbalance. Table 13 shows that balancing reduces bias towards categories and increases the reliability of detection results. Text pre-processing has a significant impact on mitigating the challenges posed by Arabic's complexity and improving the overall performance of the models, as shown in Table 13. The study utilised three primary tools: CAMeL, NLTK, and Araby. Removing stop words increased accuracy by reducing noise from unnecessary words and focusing on a vocabulary with greater semantic significance. Are these results significant? The findings of this study have confirmed the importance of data balancing and pre-processing for building a more accurate and objective model for detecting cyberbullying in the Saudi dialect. First, Data balancing mitigates the effects of class imbalance and improves the model's accuracy. Text pre-processing filters noise and focuses on the most relevant information.

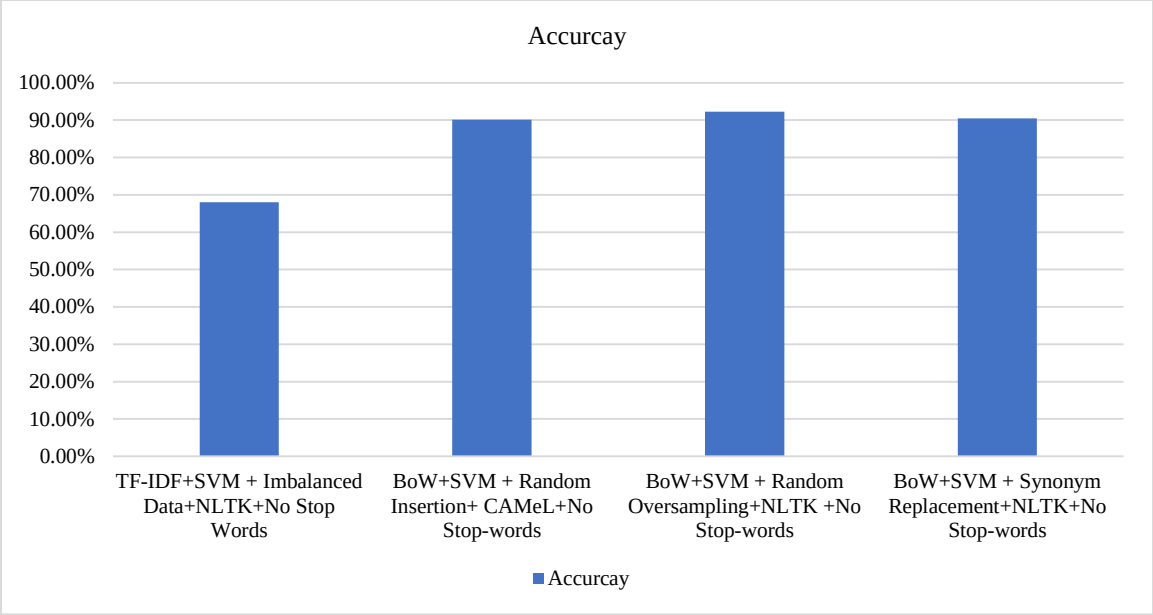

Fig. 8 Best accuracy for all scenarios

5.2.2. Statistical Significance

To confirm that the observed performance differences were not due to random variation, Wilcoxon signed-rank tests were performed across the 28 scenarios. For the best-performing model (BoW + linear SVM), all three balancing strategies, random insertion, random oversampling, and synonym replacement, produced large and statistically significant gains over the imbalanced baseline. The mean weighted F1-score increased from approximately 0.63-0.66 on the imbalanced data to around 0.86-0.90 on the balanced datasets (improvements of approximately 0.22-0.27; all $p = 0.0156$), with accuracy showing similar gains. When combined over all pre-processing and balancing conditions, linear SVM classifiers also significantly outperformed Multinomial Naïve Bayes, with mean accuracy increasing for NB from 0.79 to 0.83 for SVM and mean weighted F1-score from 0.80 to 0.83 (both $p < 0.001$). Feature extraction using BoW offered significant benefits over TF-IDF, increasing the mean accuracy from 0.81 to 0.82 and the mean weighted F1-score from 0.81 to 0.82 ($p < 0.01$). In contrast, comparisons of pipelines with and without stop word removal showed negligible changes in both accuracy and F1-score (mean differences below 0.01, all $p \geq 0.16$), and normalisation toolkits (CAMeL, NLTK, Araby) showed additional gains (0.3-0.4). Overall, these tests indicate that classifier/feature choice, class balancing, and toolkit choice have a much more substantial impact on performance.

5.2.3. Error Analysis and Misclassification Patterns

The confusion matrix shown in Figure 3 indicates that most predictions lie on the diagonal. At the same time, the remaining misclassifications are concentrated between neighbouring severity levels and in a few recurring types of errors. Table 14 presents three tweets that were incorrectly classified (paraphrased and translated to protect user privacy). In the first example, a tweet that is rated as having high severity has moral condemnation and sexual references about a friend's relationship with a girl, but the model does not predict cyberbullying. One possible reason is that the abusive content is hidden in a longer message that gives advice and mixes family issues, ethical warnings, and unclear insults, making it harder to tell the difference between abusive and non-abusive criticism. The second example is labelled as low severity and criticises a group for falsely accusing someone of "insulting" a woman; the model predicts high severity, perhaps because terms such as "people of falsehood" and "accusing" are strong lexical cues that the model associates with severe verbal attacks. The third example is a message that clearly praises the country and is not abusive, but is wrongly labelled as having low severity. This mistake could occur here because emotional or emphatic language can resemble patterns seen in abusive tweets, even when the target and intent are good. Collectively, these situations indicate that the most complex tweets involve nuanced differences in context and purpose, such as moral guidance versus insult or intense criticism versus unfounded accusation and commendation, which are difficult to identify from superficial linguistic features alone.

Table 14. Representative misclassified tweets

Tweet Paraphrased	True Label	Predicted Label
Moral condemnation of a friend's relationship with a girl, suggesting family harm and sexual motives.	High severity	Non-Cyberbullying

Criticises the group for falsely accusing someone of insulting a woman, noting his innocence and their self-interest.	Low severity	High severity
Positive message showing that national pride requires actions and results, not words.	Non-Cyberbullying	Low severity

5.2.4. Generalisability

The findings reported in this study are most relevant to Saudi dialect tweets on X (Twitter). The severity annotations, lexicon, and learned decision boundaries are shaped by the Saudi dialect and cultural norms; therefore, the system may not transfer directly to other dialects or platforms without adaptation. Nonetheless, the four-level severity framework, class-balancing approach, and experimental design can be applied in different contexts using newly annotated local data. Future work could investigate cross-dialect validation to quantify how performance changes outside the Saudi dialect Twitter domain.

5.2.5. Societal and Educational Implications:

A severity-aware detection system could act as a decision-support tool for platform administrators, school counsellors, or youth support services, helping to prioritise high-severity cases for urgent intervention while human experts review flagged content. The annotated dataset and lexicon can also guide digital citizenship programs and awareness campaigns. However, the presence of false positives and negatives shows that automated systems should complement rather than replace human judgment, with attention to privacy, transparency, and potential bias across dialects, genders, or social groups.

5.2.6. Ethical Considerations:

These ethical considerations align with discussions on responsible social-media data use. This study used public tweets gathered under Twitter's terms of service for academic research. Identifiable details were removed during pre-processing to protect privacy. As the data are publicly available and require no intervention, individual consent was not required for the study of public social media content. The analysis focused on overall patterns rather than individual users. Given potential false results, model outputs should not determine disciplinary action alone; trained moderators should review them. Privacy protection, platform compliance, and addressing biases across dialects, genders, and social groups are essential for responsible implementation.

6. Conclusion and Future Work

6.1. Conclusion

The findings show that machine learning classifiers can effectively detect four levels of cyberbullying severity when trained on balanced, pre-processed Saudi dialect tweets. Class balancing was found to be a vital step that can significantly increase the model's performance. The findings on balancing techniques, including random insertion, random oversampling, and synonym replacement, indicated that these methods reduced bias in the model's performance. In addition, training with balanced data enhanced the performance of the SVM and NB models and provided coverage and appropriate identification of the rarest categories. Pre-processing data steps were found to reduce noise and enhance important features, yielding stable improvements across all scenarios. The best accuracy was achieved with BoW + SVM using NLTK with stop-word removal, at 92.23%. In conclusion, this study has shown the successful detection of cyberbullying severity in the Saudi dialect, with 92.23% accuracy. In addition, this study found that careful balancing and pre-processing were vital for building accurate and reliable models.

6.2. Future Work

This study demonstrated the effectiveness of traditional machine learning models when combined with effective preprocessing and balancing techniques. Future work will explore deep learning approaches, such as CNNs, LSTMs, and BiLSTMs, to assess the severity of cyberbullying in Saudi dialect tweets. These models have shown promising results in natural language processing and should improve the accuracy of cyberbullying detection.

References

[1] Saudi Arabia Social Media Statistics 2024, Global Media Insight - Dubai Digital Interactive Agency, 2023. [Online]. Available: https://www.globalmediainsight.com/blog/saudi-arabia-social-media-statistics/

[2] Number of users of twitter in Saudi Arabia 2019-2028, Statista Research Department, 2025. [Online]. Available: https://www.statista.com/statistics/558404/number-of-twitter-users-in-saudi-arabia/

[3] Fadia S. AlBuhairan et al., "Time for an Adolescent health Surveillance System in Saudi Arabia: Findings from "Jeeluna"," *Journal of Adolescent Health*, vol. 57, no. 3, pp. 263-269, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[4] Monirah Abdullah Al-Ajlan, and Mourad Ykhlef, "Deep Learning Algorithm for Cyberbullying Detection," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 9, pp. 199-205, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[5] A.K. Jaithunbi et al., "Detecting Twitter Cyberbullying using Machine Learning," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 4, pp. 16307-16315, 2021. [Google Scholar] [Publisher Link]

[6] Raju Kumar, and Aruna Bhat, "A Study of Machine Learning-based Models for Detection, Control, and Mitigation of Cyberbullying in Online Social Media," *International Journal of Information Security*, vol. 21, no. 6, pp. 1409-1431, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Monirah A. Al-Ajlan, and Mourad Ykhlef, "Optimized Twitter Cyberbullying Detection based on Deep Learning," *2018 21st Saudi Computer Society National Computer Conference (NCC)*, Riyadh, Saudi Arabia, pp. 1-5, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[8] Alanoud Mohammed Alduailaj, and Aymen Belghith, "Detecting Arabic Cyberbullying Tweets Using Machine Learning," *Machine Learning and Knowledge Extraction*, vol. 5, no. 1, pp. 29-42, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Djedjiga Mouheb et al., "Detection of Arabic Cyberbullying on Social Networks using Machine Learning," *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, Abu Dhabi, United Arab Emirates, pp. 1-5, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[10] Deema Alghamdi et al., "Automatic Detection of Cyberbullying and Threatening in Saudi Tweets using Machine Learning," *International Journal of Advanced and Applied Sciences*, vol. 8, no. 10, pp. 17-25, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Sourabh Parime, and Vaibhav Suri, "Cyberbullying Detection and Prevention: Data Mining and Psychological Perspective," *2014 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2014]*, Nagercoil, India, pp. 1541-1547, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[12] Marilyn Campbell, and Sheri Bauman, *Cyberbullying: Definition, Consequences, Prevalence*, Reducing Cyberbullying in Schools: International Evidence-based Best Practices, Academic Press, pp. 3-16, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[13] Vikas S. Chavan, and S.S. Shylaja, "Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social Media Network," *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Kochi, India, pp. 2354-2358, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[14] Andreas König, Mario Gollwitzer, and Georges Steffgen, "Cyberbullying as an Act of Revenge?," *Journal of Psychologists and Counsellors in Schools*, vol. 20, no. 2, pp. 210-224, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[15] Peter K. Smith et al., "Cyberbullying: Its Nature and Impact in Secondary School Pupils," *Journal of Child Psychology and Psychiatry*, vol. 49, no. 4, pp. 376-385, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[16] Sydney L. Brunecz, "More Harm than Good? Why Schools Who Take a Zero-Tolerance Stance on Cyberbullying Cause More Problems than Solutions," *Case Western Reserve Journal of Law, Technology & the Internet*, vol. 6, no. 1, pp. 13-42, 2014. [Google Scholar] [Publisher Link]

[17] Allison Paolini, "Cyberbullying: Role of the School Counselor in Mitigating the Silent Killer Epidemic," *International Journal of Educational Technology*, vol. 5, no. 1, pp. 1-8, 2018. [Google Scholar] [Publisher Link]

[18] Ye Zhang Pogue, The Digital Dagger: The Destructive Impact of Cyberbullying, Psychology Today, 2023. [Online]. Available: https://www.psychologytoday.com/us/blog/the-human-identity/202307/the-digital-dagger-the-destructive-impact-of-cyberbullying?msockid=2f3d6626b97162203f4d74a4bd716cea

[19] Ditch the Label, Cyberbullying Statistics: What They Tell Us, Ditch the Label Youth Charity, 2017. [Online]. Available: https://www.ditchthelabel.org/cyber-bullying-statistics-what-they-tell-us

[20] Deborah Goebert et al., "The Impact of Cyberbullying on Substance Use and Mental Health in A Multiethnic Sample," *Maternal and Child Health Journal*, vol. 15, no. 8, pp. 1282-1286, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[21] Tanya Beran, and Qing Li, "The Relationship between Cyberbullying and School Bullying," *The Journal of Student Wellbeing*, vol. 1, no. 2, pp. 16-33, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[22] Justin W. Patchin, Sameer Hinduja, Summary of Our Cyberbullying Research (2007-2025), Cyberbullying Research Center, 2024. [Online]. Available: https://cyberbullying.org/summary-of-our-cyberbullying-research

[23] Victoria Brown, Elizabeth Clery, and Christopher Ferguson, "*Estimating the Prevalence of Young People Absent from School Due to Bullying*," National Centre for Social Research, 2011. [Google Scholar]

[24] Ainoa Mateu et al., "Cyberbullying and Post-Traumatic Stress Symptoms in UK Adolescents," *Archives of Disease in Childhood*, vol. 105, no. 10, pp. 951-956, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[25] Njoud Alrasheed et al., "Prevalence and Risk Factors of Cyberbullying and its Association with Mental Health among Adolescents in Saudi Arabia," *Cureus*, vol. 14, no. 12, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[26] Gassem Gohal et al., "Prevalence and Related Risks of Cyberbullying and its Effects on Adolescent," *BMC psychiatry*, vol. 23, no. 1, pp 1-10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[27] Nawal A. Alissa, and Rawan Abu Shryei, "Cyberbullying among Female College Students in Saudi Arabia," *International Journal of Child, Youth and Family Studies*, vol. 16, no. 1, pp. 52-66, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[28] Damian Maher, "Cyberbullying: An Ethnographic Case Study of One Australian Upper Primary School Class," *Youth Studies Australia*, vol. 27, no. 4, pp. 50-57, 2008. [Google Scholar] [Publisher Link]

[29] Batoul Haidar, Maroun Chamoun, and Fadi Yamout, "Cyberbullying Detection: A Survey on Multilingual Techniques," *2016 European Modelling Symposium (EMS)*, Pisa, Italy, pp. 165-171, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[30] Samaneh Nadali et al., "A Review of Cyberbullying Detection: An Overview," *2013 13th International Conference on Intellient Systems Design and Applications*, Salangor, Malaysia, pp. 325-330, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[31] Norulzahrah Mohd Zainudin et al., "A Review on Cyberbullying in Malaysia from Digital Forensic Perspective," *2016 International Conference on Information and Communication Technology (ICICTM)*, Kuala Lumpur, Malaysia, pp. 246-250, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[32] Nancy E. Willard, *Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress*, Research press, 2025. [Google Scholar] [Publisher Link]

[33] Jennifer Bayzick, April Kontostathis, and Lynne Edwards, "Detecting the Presence of Cyberbullying using Computer Software," *WebSci Conference*, Koblenz, Germany, pp. 1-2, 2011. [Google Scholar]

[34] Taeho Jo, *Machine Learning Foundations*, Supervised, Unsupervised, and Advanced Learning, Springer Cham, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[35] Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana, "Cybercrime Detection in Online Communications: The Experimental Case of Cyberbullying Detection in the Twitter Network," *Computers in Human Behavior*, vol. 63, pp. 433-443, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[36] Michele Di Capua, Emanuel Di Nardo, and Alfredo Petrosino, "Unsupervised Cyber Bullying Detection in Social Networks," *2016 23rd International Conference on Pattern Recognition (ICPR)*, Cancun, Mexico, pp. 432-437, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[37] Xiaowei Gu, "A Self-Training Hierarchical Prototype-based Approach for Semi-Supervised Classification," *Information Sciences*, vol. 535, pp. 204-224, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[38] Vinita Nahar et al., "Semi-Supervised Learning for Cyberbullying Detection in Social Networks," *Databases Theory and Applications: 25th Australasian Database Conference*, Brisbane, QLD, Australia, pp. 160-171, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[39] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[40] Shervin Minaee et al., "Deep Learning--based Text Classification: A Comprehensive Review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1-40, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[41] Celestine Iwendi et al., "Cyberbullying Detection Solutions based on Deep Learning Architectures," *Multimedia Systems*, vol. 29, no. 3, pp. 1839-1852, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[42] K.G. Apoorva, and D. Uma, "Detection of Cyberbullying Using Machine Learning and Deep Learning Algorithms," *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, Ravet, India, pp. 1-7, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[43] Jalal Omer Atoum, "Cyberbullying Detection Neural Networks using Sentiment Analysis," *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA, pp. 158-164, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[44] Roman Egger, and Enes Gokce, *Natural Language Processing (NLP): An Introduction: Making Sense of Textual Data*, Applied Data Science in Tourism, Springer, Cham, pp. 307-334, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[45] K.R. Chowdhary, *Natural Language Processing*, Fundamentals of Artificial Intelligence, Springer, New Delhi, pp. 603-649, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[46] Dipanjan Sarkar, *Text Analytics with Python*, A Practitioner's Guide to Natural Language Processing, Apress Berkeley, CA, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[47] Elizabeth D. Liddy, Natural Language Processing, 2nd Ed., Encyclopedia of Library and Information Science, NY, Marcel Decker, Inc, 2001. [Google Scholar] [Publisher Link]

[48] Yue Kang et al., "Natural Language Processing (NLP) in Management Research: A Literature Review," *Journal of Management Analytics*, vol. 7, no. 2, pp. 139-172, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[49] Muhammad Abdul-Mageed, and Mona Diab, "AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis," *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, vol. 515, pp. 3907-3914, 2012. [Google Scholar] [Publisher Link]

[50] Hossam S. Ibrahim, Sherif M. Abdou, and Mervat Gheith, "Sentiment Analysis for Modern Standard Arabic and Colloquial," *arXiv Preprint*, pp. 95-109, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[51] Kenneth R. Beesley, "Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001," *ACL Workshop on Arabic Language Processing: Status and Perspective*, vol. 1, pp. 1-8, 2001. [Google Scholar]

[52] Tim Buckwalter, "Issues in Arabic Orthography and Morphology Analysis," *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, Switzerland, pp. 31-34, 2004. [Google Scholar] [Publisher Link]

[53] Mohamed Elmahdy et al., "Survey on Common Arabic Language Forms from a Speech Recognition Point of View," *Proceeding of International Conference on Acoustics (NAG-DAGA)*, Rotterdam, pp. 63-66, 2009. [Google Scholar] [Publisher Link]

[54] Kareem Darwish et al., "A Panoramic Survey of Natural Language Processing in the Arab World," *Communications of the ACM*, vol. 64, no. 4, pp. 72-81, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[55] Mohamed Abd Elaziz et al., *Recent Advances in NLP: The Case of Arabic Language*, Springer Cham, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[56] Batoul Haidar, Maroun Chamoun, and Ahmed Serhrouchni, "Multilingual Cyberbullying Detection System: Detecting Cyberbullying in Arabic Content," *2017 1st cyber security in networking conference (CSNet)*, Rio de Janeiro, Brazil, pp. 1-8, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[57] Azalden Alakrot, Liam Murray, and Nikola S. Nikolov, "Towards Accurate Detection of Offensive Language in Online Communication in Arabic," *Procedia Computer Science*, vol. 142, pp. 315-320, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[58] Dhiaa Musleh et al., "A Machine Learning Approach to Cyberbullying Detection in Arabic Tweets," *Computers, Materials & Continua*, vol. 80, no. 1, pp. 1033-1054, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[59] Bandeh Ali Talpur, and Declan O'Sullivan, "Multi-Class Imbalance in Text Classification: A Feature Engineering Approach to Detect Cyberbullying in Twitter," *Informatics*, vol. 7, no. 4, pp. 1-22, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[60] M. Rahman, S. Nur, M. T. Ahmed, D. Das, and A. T. Islam, "A Feature Engineering Approach for Detecting Cyberbullying in Bangla Text using Machine Learning," *2022 International Conference on Recent Progresses in Science, Engineering and Technology (ICRPSET)*, Rajshahi, Bangladesh, pp. 1-5, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[61] Jheng-Long Wu, and Chiao-Yu Tang, "Classifying The Severity of Cyberbullying Incidents by using A Hierarchical Squashing-Attention Network," *Applied Sciences*, vol. 12, no. 7, pp. 1-19, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[62] Madhura Vikram Vyawahare, and Sharvari Govilkar, "Severity Detection of Cyberbullying in Online Social Networks Using Machine Learning," *2022 5th International Conference on Advances in Science and Technology (ICAST)*, Mumbai, India, pp. 1-6, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[63] Sylvia W. Azumah et al., "Cyberbullying in Text Content Detection: An Analytical Review," *International Journal of Computers and Applications*, vol. 45, no. 9, pp. 579-586, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[64] Tanjim Mahmud et al., "Cyberbullying Detection for Low-Resource Languages and Dialects: Review of the State of the Art," *Information Processing & Management*, vol. 60, no. 5, pp 1-52, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[65] Hooayda Allwaibed et al., "Cyberbullying Detection Approaches for Arabic Texts: Systematic Literature Review," *Frontiers in Artificial Intelligence*, vol. 8, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[66] Bandeh Ali Talpur, and Declan O'Sullivan, "Cyberbullying Severity Detection: A Machine Learning Approach," *PloS One*, vol. 15, no. 10, pp. 1-19, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[67] Jason Wei, and Kai Zou, "Eda: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 6382-6388, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[68] Anna Glazkova, "A Comparison of Synthetic Oversampling Methods for Multi-Class Text Classification," *arXiv Preprint*, pp. 1-12, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[69] Ossama Obeid et al., "CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing," *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, pp. 7022-7032, 2020. [Google Scholar] [Publisher Link]

[70] Edward Loper, and Steven Bird, "Nltk: The Natural Language Toolkit," *arXiv Preprint*, pp. 1-8, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[71] Taha Zerrouki, "PyArabic: A Python Package for Arabic Text," *Journal of Open Source Software*, vol. 8, no. 84, pp. 1-6, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[72] Corinna Cortes, and Vladimir Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995. [CrossRef] [Google Scholar] [Publisher Link]

[73] David M.W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," *arXiv Preprint*, pp. 37-63, 2020. [CrossRef] [Google Scholar] [Publisher Link]

Table 13. Outcomes of 28 experimental scenarios

Scenario	BoW+SVM				BoW+NB				TF-IDF+SVM				TF-IDF+NB			
	Accuracy	F1-Score	Recall	Precision	Accuracy	F1-Score	Recall	Precision	Accuracy	F1-Score	Recall	Precision	Accuracy	F1-Score	Recall	Precision
Original Data - No preprocessing	64.80%	0.6209	0.6480	0.6221	62.48%	0.6288	0.6248	0.6428	63.43%	0.6224	0.6343	0.6155	64.72%	0.6321	0.6472	0.6253
Original Data+ CAMeL	65.32%	0.6294	0.6532	0.6261	61.45%	0.6178	0.6145	0.6326	64.63%	0.6327	0.6463	0.6250	65.83%	0.6325	0.6583	0.6210
Original Data+ CAMeL+ stop words removal	66.52%	0.6361	0.6652	0.6496	62.65%	0.6298	0.6265	0.6438	67.99%	0.6544	0.6799	0.6488	66.18%	0.6384	0.6618	0.6274
Original Data+ NLTK	64.97%	0.6255	0.6497	0.6221	61.19%	0.6152	0.6119	0.6296	64.11%	0.6288	0.6411	0.6230	65.40%	0.6288	0.6540	0.6175
Original Data+ NLTK+ stop words removal	66.78%	0.6361	0.6678	0.6553	62.31%	0.6265	0.6231	0.6390	67.47%	0.6493	0.6747	0.6419	66.70%	0.6442	0.6670	0.6331
Original Data+ Araby	64.63%	0.6196	0.6463	0.6175	63.43%	0.6361	0.6343	0.6515	63.17%	0.6201	0.6317	0.6141	66.44%	0.6366	0.6644	0.6285
Original Data+ Araby+ stop words removal	65.15%	0.6200	0.6515	0.6420	63.86%	0.6412	0.6386	0.6558	66.61%	0.6409	0.6661	0.6313	66.27%	0.6392	0.6627	0.6300
Random Insertion balanced Data- No perprocessing	89.00%	0.8901	0.8900	0.8927	83.87%	0.8358	0.8387	0.8466	86.50%	0.8639	0.8650	0.8711	82.52%	0.8196	0.8252	0.8410
Random Insertion balanced Data+CAMeL	89.32%	0.8934	0.8932	0.8967	84.19%	0.8393	0.8419	0.8484	86.04%	0.8595	0.8604	0.8667	83.36%	0.8297	0.8336	0.8461
Random Insertion balanced Data+CAMeL+ stop words removal	90.11%	0.9012	0.9011	0.9072	84.28%	0.8402	0.8428	0.8502	88.21%	0.8819	0.8821	0.8836	83.26%	0.8326	0.8326	0.8432
Random Insertion balanced Data+NLTK	89.00%	0.8900	0.8900	0.8931	84.14%	0.8388	0.8414	0.8483	86.41%	0.8633	0.8641	0.8707	83.22%	0.8322	0.8322	0.8450
Random Insertion balanced	90.06%	0.9005	0.9006	0.9052	84.14%	0.8386	0.8414	0.8494	88.49%	0.8848	0.8849	0.8863	83.36%	0.8297	0.8336	0.8446

Data+NLTK+ stop words removal																
Random Insertion balanced Data+Araby	89.04%	0.8905	0.8904	0.8932	84.14%	0.8384	0.8414	0.8485	86.50%	0.8639	0.8650	0.8710	83.03%	0.8248	0.8303	0.8434
Random Insertion balanced Data+Araby+ stop words removal	90.01%	0.9002	0.9001	0.9058	84.23%	0.8392	0.8423	0.8493	88.44%	0.8844	0.8844	0.8859	83.73%	0.8333	0.8373	0.8474
Random Oversampling balanced Data-No preprocessing	91.49%	0.9150	0.9149	0.9156	85.76%	0.8542	0.8576	0.8622	88.77%	0.8871	0.8877	0.8940	84.74%	0.8422	0.8474	0.8586
Random Oversampling balanced Data+CAMeL	91.35%	0.9138	0.9135	0.9149	86.41%	0.8611	0.8641	0.8692	88.90%	0.8882	0.8890	0.8945	85.11%	0.8463	0.8511	0.8616
Random Oversampling balanced Data+CAMeL+ stop words removal	91.86%	0.9188	0.9186	0.9215	85.53%	0.8517	0.8553	0.8603	89.88%	0.8987	0.8988	0.8995	85.34%	0.8491	0.8534	0.8615
Random Oversampling balanced Data+NLTK	91.17%	0.9119	0.9117	0.9130	86.27%	0.8593	0.8627	0.8680	88.49%	0.8843	0.8849	0.8902	85.16%	0.8466	0.8516	0.8620
Random Oversampling balanced Data+NLTK+ stop words removal	92.23%	0.9225	0.9223	0.9251	85.48%	0.8512	0.8548	0.8601	90.43%	0.9043	0.9043	0.9049	85.34%	0.8491	0.8534	0.8607
Random Oversampling balanced Data+Araby	91.22%	0.9123	0.9122	0.9131	85.81%	0.8547	0.8581	0.8622	88.58%	0.8850	0.8858	0.8930	85.21%	0.8474	0.8521	0.8619

Random Oversampling balanced Data+Araby+ stop words removal	92.22%	0.9201	0.9200	0.9230	85.48%	0.8512	0.8548	0.8603	89.92%	0.8990	0.8992	0.9000	84.79%	0.8434	0.8479	0.8561
Synonym Replacement balanced Data-No preprocessing	88.95%	0.8895	0.8895	0.8910	84.95%	0.8408	0.8442	0.8512	86.73%	0.8665	0.8673	0.8723	83.36%	0.8278	0.8336	0.8467
Synonym Replacement balanced Data+CAMeL	89.51%	0.8952	0.8951	0.8972	84.84%	0.8452	0.8484	0.8560	87.15%	0.8704	0.8715	0.8778	84.14%	0.8367	0.8414	0.8527
Synonym Replacement balanced Data+CAMeL+ stop words removal	89.78%	0.8980	0.8978	0.9028	84.84%	0.8448	0.8484	0.8554	88.07%	0.8806	0.8807	0.8819	83.91%	0.8345	0.8391	0.8489
Synonym Replacement balanced Data+NLTK	89.27%	0.8929	0.8927	0.8951	84.74%	0.8443	0.8474	0.8548	87.15%	0.8706	0.8715	0.8768	83.96%	0.8346	0.8396	0.8508
Synonym Replacement balanced Data+NLTK+ stop words removal	90.43%	0.9044	0.9043	0.9082	84.79%	0.8447	0.8479	0.8552	88.86%	0.8884	0.8886	0.8894	84.00%	0.8358	0.8400	0.8507
Synonym Replacement balanced Data+Araby	88.95%	0.8896	0.8895	0.8912	84.79%	0.8444	0.8479	0.8543	86.73%	0.8663	0.8673	0.8726	83.26%	0.8272	0.8326	0.8431
Synonym Replacement balanced Data+Araby+ stop words removal	89.78%	0.8980	0.8987	0.9028	84.84%	0.8448	0.8484	0.8554	88.07%	0.8806	0.8807	0.8819	83.91%	0.8345	0.8391	0.8489