



Detecting Arabic Cyberbullying Tweets in Arabic Social Using Deep Learning

الكشف عن تغريدات التنمر الإلكتروني باللغة العربية على مواقع التواصل
الاجتماعي العربية باستخدام التعلم العميق

by

FARIS KHAMIS ATIQ ALFALASI

Dissertation submitted in partial fulfilment

of the requirements for the degree of

MSc INFORMATICS

at

The British University in Dubai

June 2023

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the contents of this dissertation for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

A handwritten signature in black ink, appearing to read 'Faris', written in a cursive style.

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

The author whose copyright is declared on the title page of the work has granted to the British University in Dubai the right to lend his/her research work to users of its library and as an open source and to make partial or single copies for educational and research use.

The author has also granted permission to the University to keep or make a digital copy for similar use and for the purpose of preservation of the work digitally.

Multiple copying of this work for scholarly purposes may be granted by either the author, the Registrar or the Dean only.

Copying for financial gain shall only be allowed with the author's express permission.

Any use of this work in whole or in part shall respect the moral rights of the author to be acknowledged and to reflect in good faith and without detriment the meaning of the content, and the original authorship.

ABSTRACT

The widespread engagement with social media platforms in recent years has made cyberbullying a significant concern. Individuals may have catastrophic side effects from that as well, including despair, anxiety, and even suicide. Due to the difficulty of manually detecting and categorizing vast volumes of electronic text data, conventional methods for recognizing and combating cyberbullying have not proven successful. As a consequence, deep learning methods have become a potential solution for this situation. Artificial neural networks and other deep learning approaches can automatically identify patterns and features from a massive quantity of data. These methods may be applied to electronic text data analysis to spot cyberbullying-related trends. Techniques for natural language processing may be used to text data to extract useful features like sentiment, emotion, and subjectivity. A sizable dataset of electronic text data was gathered from multiple social media platforms like Twitter, Instagram, YouTube, and many more sites in order to examine cyberbullying in social media using machine learning and deep learning techniques. The data needs to be initially prepared so that deep learning algorithms may be trained on it before cyberbullying analysis can be done. Manually annotated data from a corpus collection was used to label the information for deep learning purposes. Pre-processing is a vital part of the data preparation process for cyberbullying detection. There are several varieties of Arabic, but the three most common are dialect Arabic, Modern Standard Arabic, and Classical Arabic. Because of its widespread use on social media, DA Arabic is the subject of this essay. Based on the existence of cyberbullying, the data was then preprocessed and classified. In this work, two cases of classification were adapted. The first case was 2-classes classification where the data labeled as either cyberbullying or not cyberbullying. The second case was 6-classes classification which consists of six different cyberbullying types. To categorize electronic text in these two cases, deep learning models such as convolutional neural networks and recurrent neural networks and a combination of CNN-RNN were trained on this data. In an independent test set, the trained models were assessed, and they showed promise in identifying cyberbullying via social media. The results that obtained from 2-classes classification showed a superiority of LSTM in terms of accuracy with 95.59%, while the best accuracy in the 6-classes classification gained from implementing CNN with 78.75%. Meanwhile the f1-score results were the highest in LSTM for the 2-lasses and 6-classes classifications with 96.73%, and 89%, respectively. These findings emphasize the potential for deep learning techniques to be applied in the development of

automated systems for identifying and combating cyberbullying on social media and show how well they work in detecting cyberbullying.

ABSTRACT (in Arabic)

أدى التفاعل الواسع النطاق مع منصات وسائل التواصل الاجتماعي في السنوات الأخيرة إلى جعل التتبع عبر الإنترنت مصدر قلق كبير. قد يكون للأفراد آثار جانبية كارثية من ذلك أيضًا ، بما في ذلك اليأس والقلق وحتى الانتحار. نظرًا لصعوبة اكتشاف وتصنيف كميات هائلة من البيانات النصية الإلكترونية يدويًا ، فإن الطرق التقليدية للتعرف على التتبع الإلكتروني ومكافحته لم تثبت نجاحها. نتيجة لذلك ، أصبحت أساليب التعلم العميق حلاً محتملاً لهذا الموقف. يمكن للشبكات العصبية الاصطناعية وغيرها من مناهج التعلم العميق تحديد الأنماط والميزات تلقائيًا من كمية هائلة من البيانات. يمكن تطبيق هذه الأساليب على تحليل البيانات النصية الإلكترونية لاكتشاف الاتجاهات المتعلقة بالتسلط عبر الإنترنت. يمكن استخدام تقنيات معالجة اللغة الطبيعية لكتابة البيانات النصية لاستخراج ميزات مفيدة مثل المشاعر والعاطفة والذاتية. تم جمع مجموعة بيانات كبيرة من البيانات النصية الإلكترونية من العديد من منصات الوسائط الاجتماعية مثل Twitter و Instagram و YouTube والعديد من المواقع الأخرى من أجل فحص التتبع عبر الإنترنت في وسائل التواصل الاجتماعي باستخدام التعلم الآلي وتقنيات التعلم العميق. يجب إعداد البيانات مبدئيًا بحيث يمكن تدريب خوارزميات التعلم العميق عليها قبل إجراء تحليل التتبع الإلكتروني. تم استخدام البيانات المشروحة يدويًا من مجموعة المدونات لتسمية المعلومات لأغراض التعلم العميق. تعد المعالجة المسبقة جزءًا حيويًا من عملية إعداد البيانات لاكتشاف التسلط عبر الإنترنت. هناك عدة أنواع من اللغة العربية ، ولكن الأنواع الثلاثة الأكثر شيوعًا هي اللهجة العربية والعربية الفصحى الحديثة والعربية الفصحى. نظرًا لاستخدامها على نطاق واسع على وسائل التواصل الاجتماعي ، فإن DA Arabic هي موضوع هذا المقال. بناءً على وجود التسلط عبر الإنترنت ، تمت معالجة البيانات وتصنيفها مسبقًا. في هذا العمل ، تم تكيف حالتين من التصنيف. كانت الحالة الأولى عبارة عن تصنيف من فئتين حيث تم تصنيف البيانات على أنها إما تسلط عبر الإنترنت أو ليس تسلطًا عبر الإنترنت. كانت الحالة الثانية عبارة عن تصنيف من 6 فئات تتكون من ستة أنواع مختلفة من التسلط عبر الإنترنت. لتصنيف النص الإلكتروني في هاتين الحالتين ، تم تدريب نماذج التعلم العميق مثل الشبكات العصبية التلافيفية والشبكات العصبية المتكررة ومجموعة من CNN-RNN على هذه البيانات. في مجموعة اختبار مستقلة ، تم تقييم النماذج المدربة ، وأظهرت نتائج واعدة في تحديد التتبع عبر الإنترنت عبر وسائل التواصل الاجتماعي. أظهرت النتائج التي تم الحصول عليها من تصنيف فئتين تفوق LSTM من حيث الدقة بنسبة 95.59% ، بينما حصلت أفضل دقة في تصنيف 6 فئات من تطبيق CNN بنسبة 78.75%. وفي الوقت نفسه ، كانت نتائج F1 هي الأعلى في LSTM لتصنيفي 2-classes و 6-class بنسبة 96.73% و 89% على التوالي. تؤكد هذه النتائج على إمكانية تطبيق تقنيات التعلم العميق في تطوير الأنظمة الآلية لتحديد ومكافحة التسلط عبر الإنترنت على وسائل التواصل الاجتماعي وإظهار مدى نجاحها في اكتشاف التتبع الإلكتروني.

ACKNOWLEDGEMENTS

I would like to acknowledge the whole community at British University in Dubai (BUID) for providing me the opportunity to explore and major in informatics, which has given me a wealth of information.

I also want to thank Prof. Manar Alkhatib for supervising this dissertation and creating it possible for me to complete this work. I also like to thank Dr. Khaled Shaalan for his assistance and counsel during this study. It would have been tough to generate a research paper of this caliber without their support.

Lastly, I would want to express my gratitude to my family, particularly to my sisters, and parents, who all encouraged and supported me in pursuing my education.

TABLE OF CONTENTS

1. CHAPTER ONE: Introduction	1
1.1 Problem Definition	2
1.2 Research Objectives	3
1.3 Thesis Contribution	3
1.4 Dissertation Organization	3
2. CHAPTER TWO: Related Work	5
2.1 Arabic Sentiment Analysis Using Machine Learning and Neural Network	5
2.2 Arabic Cyberbullying Using Machine Learning and Neural Network	8
2.3 English Cyberbullying Using Machine Learning and Neural Network	10
3. CHAPTER THREE: Literature review	12
3.1 Background	12
3.1.1 Cyberbullying Definition	12
3.1.2 Cyberbullying on Social Media	13
3.1.3 Cyberbullying Forms	13
3.1.4 Sentiment Analysis	15
3.1.5 Emotion	15
3.1.6 Arabic Language	16
3.1.7 Arabic Text Cyberbullying Types	18
3.2 Deep Learning Approaches	20
3.2.1 Convolutional neural networks	20
3.2.2 Recurrent Neural Networks (RNN)	23
3.2.3 Long Short Term Memory Networks (LSTM)	25
3.2.4 Hybrid CNN-RNN	26
3.2.5 A comparison between deep learning models	27
3.3 Data Pre-Processing	29
3.3.1 Data Cleaning	29
3.3.2 Vectorization and Word Embedding	31
3.3.3 Continuous Bag-of-Words Model (CBOW)	32
3.3.4 Skip-Gram (SG) model	33
3.3.5 Feature Extraction	34
3.3.6 Annotation	35
3.3.7 Training Data And Test Data	36
3.3.8 Metrics Evaluation	37
4. CHAPTER FOUR: Dataset Corpus	39
4.1 Data Collection	39
4.2 Two-Class Classification	40

4.3	Six-Class Classification	40
5	CHAPTER FIVE: Methodology	42
5.1	Data Preprocessing Methodology	43
5.1.2	Data Preparation	43
5.1.1	Collecting Data	44
5.1.3	Training dataset and test data	44
5.2	Feature Extraction	47
5.3	Deep learning Models	47
5.4	Evaluation metrics	47
6	CHAPTER SIX: Evaluation and Results	50
6.1	Experiment Setup	50
6.1.1	Data Preprocessing	50
6.1.2	Deep Learning Neural Network Models	50
6.1.2.1	Feature Representation	50
6.1.2.2	CNN Architecture	51
6.1.2.3	Evaluation metrics	53
6.1.2.4	LSTM Architecture	53
6.2	Experiments Results	54
6.2.1	Results and Discussion	54
7	CHAPTER SEVEN: Conclusion And Future Work	61
	References	63

LIST OF TABLES

Table 1: Cyberbullying Types and examples	20
Table 2 : Characteristics Of The Two-Class Classification.	40
Table 3: An Example Of The Collected Data For The First Dataset	40
Table 4: Characteristics of the Six-Class Classification	41
Table 5: An Example Of The Collected Data For The Six-Class Classification	41
Table 6: Training Data and Test Data	45
Table 7: Training Data And Test Data In Two-Class and Six-Class Classifications.....	45
Table 8: Summarize The Hyper-Parameters Of The CNN Model	52
Table 9: hyper-parameters for the LSTM model.....	53
Table 10: Hyper-parameters for CNN-LSTM.....	53
Table 11: performance for the 2-classes classification	55
Table 12: Results Of CNN On The 6-Classes Classification	56
Table 13: Performance Of LSTM On 6 Classes Classification.....	57
Table 14: Performance Of The CNN-LSTM on the 6 Classes Classification.....	59

LIST OF FIGURES

Figure 1: Cyberbullying Forms	14
Figure 2: CNN Convolutional Layer.....	21
Figure 3: CNN Pooling Layer	22
Figure 4: CNN Architecture.....	22
Figure 5: Visualizing in CNN	23
Figure 6: CNN Architecture.....	23
Figure 7: Basic architecture of Recurrent Neural Networks	25
Figure 8: LSTM Single Cell.....	26
Figure 9: CNN-LSTM Architecture	27
Figure 10: Data Preprocessing and Cleaning	30
Figure 11: Raw data crawled from Tweeter	39
Figure 12: Methodology Structure	43
Figure 13: 2- Classes Classification and 6-clses classification Datasets Related Statistics	45
Figure 14: the data set for the 2-Class Classification.....	46
Figure 15: the data set for the 6-Class Classification.....	46
Figure 16: Training data categories for the 6-Class Classification	46
Figure 17: Test Data categories in 6-Classes Classification	46
Figure 18: dataset categories for the 6-Class Classification	46
Figure 19: LSTM Architecture.....	49
Figure 20: CNN Architecture	49
Figure 21: CNN-LSTM Architecture	49
Figure 22: Convolutional neural network (CNN) architecture.	52
Figure 23: Results of 2 classes classification	54
Figure 24: Accuracy on 2 Classes Classification	55
Figure 25: Results of implementing CNN on 6 classes classification	57
Figure 26: Results of LSTM on 6 Classes Classification.....	58
Figure 27: Results of CNN-LSTM on 6 Classes Classification.....	58
Figure 28: the Accuracy in 6 Classes Classification	59
Figure 29: Accuracy in both 2 Classes and 6- Classes classifications	60

LIST OF ABBREVIATIONS

Term	Abbreviations
ACB	Animal Cyberbullying
CA	Classic Arabic
CBOW	Continuous Bag-of-Words Model
CNN	Convolutional Neural Network
DA	Dialect Arabic
DT	Decision Tree
DL	Deep Learning
KNN	K-Nearest Neighbor
LR	Logistic Regression
LCB	Looking Cyberbullying
LSTM	LSTM
ML	Machine Learning
MLP	multilayer perceptron
MSA	Modern Standard Arabic
NB	Naive Bayes
NCB	Not Cyberbullying
PSCB	Psychological Cyberbullying
RF	Random Forest
RCB	Religious Cyberbullying
RNN	Recurrent Neural Networks
SA	Sentiment Analysis
SCB	Sexual Cyberbullying
SG	Skip-Gram Model
SVM	Support Vector Machines

1. CHAPTER ONE: INTRODUCTION

Nowadays, the Advances in technology have greatly improved our ability to access information, communicate with others, and complete tasks quickly and efficiently. However, there are also concerns about the negative impacts of technology, such as its effects on cyber-attacks and threats to privacy, security, and social interactions. Statistics declared that about 18% of kids in Europe were suffered either by people bullying or effected via harassing them by the Internet connections and mobile communication. Twitter is a common interacting social networking platform that permits users to share their information by tweets or comments.

All the People around the world now are technologically communicated, regardless of timing, the location or the distance between them. Social networks platforms are trending nowadays, and the majority of the people, particularly the teenagers, are passionate to join and interconnect with the online relationships. The anonymity of social media, where users usually utilize fake names rather than their real names, has resulted in a massive number of cybercrimes. There are several types of online crime, for instance, cyberbullying which cause suffering for the bullied.

Cyberbullying is considered a severe ethical affair on the internet platforms, and the number of bullied persons who have been cyberbullying victims, mostly teenagers, is annoying. Cyberbullying refers to the use of technology, such as social media, to harass, intimidate, or otherwise harm others. This can include actions such as posting hurtful comments or messages, sharing embarrassing pictures or videos, or creating fake profiles to mock or harass someone. Cyberbullying can be particularly harmful because it can happen 24/7 and can be difficult for the victim to escape from, as social media is often a constant presence in their lives. Additionally, the anonymity and distance provided by the internet can make it easier for bullies to act aggressively and without consequence.

In order to estimate cyberbullying propagation, various research papers have discussed cyberbullying, and the results proved that cyberbullying is a common issue between the youth nowadays, with an increasing number of victims [1]. Various techniques of cyberbullying detection have been structured to monitor and prevent cyberbullying.

Studies have increased in the domain of cyberbullying detection. In spite of the propagation and the negative impact within the Arabic culture, some researchers have

studied this form of offensive in the Arabic language [2]. The cyberbullying on the Arabic platforms is yet restricted and is considered a problematic work due to several causes: First, the Arabic spoken language has extremely complicated morphological entities. Second, the majority of the Arabic people speak colloquial Arabic rather than modern standard Arabic (MSA). Actually, there are various words that are not permitted in the Arabic culture, meanwhile, they are completely agreeable in other societies [3]. For instance, the expression “dog” “كلب” or the word “حمار” “donkey” are examples of farm animals. Nevertheless, it is not allowed to utilize these forms of expressions in different contexts, for example describing people or actions. Nowadays, the cyberbullying automatic detection has led to significant development in cyberbullying classifications, particularly in English language, like in [4] and [5]. However, few studies have been done utilizing deep learning for Arabic cyberbullying detection on social media thus, we aim in this study to enhance the accuracy of the Arabic cyberbullying detection and improve the performance of the system.

1.1 Problem Definition

Cyberbullying detection is one of the essential domains that is widely extended due to the increase in online communications. The definition of cyberbullying is “willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices”, is to identify the reviewer’s feelings toward a topic using textual context. The goal of Cyberbullying detection is to identify the cyberbullying on social media using the Arabic language in the textual context.

The problem statement can be depicted as the detecting reviews polarity to be classified as negative, positive, or neutral and specify the negative polarity as cyberbullying. The reality that the Arabic language has not been considered as much as the English language is a wake-up call for all Arabic scientific researchers. Nevertheless, in the past few years, we can notice a remarkable effort done in Arabic cyberbullying detection resources. The Arabic language has much more challenges than foreign languages due to its complex morphology and the presence of different dialects.

This study additionally examines how significant it is to interest from deep learning for detecting cyberbullying on a large scale such as in social media using Arabic language. The research challenge is evaluated by contrasting the performance of many deep learning classifiers with a number of datasets. The outcomes will assist us recognize how helpful deep learning models are in classifying Arabic text and detecting cyberbullying in social media.

1.2 Research Objectives

The main goal of this research is to seek for answering to the following challenges

- To classify the Arabic tweets into two categories cyberbullying and Non-cyberbullying.
- To generate a table of “Arabic cyberbullying” expressions from the collected data.
- To apply deep learning classifiers on the collected dataset and compare the empirical results with conventional learning techniques results obtained from previous research.
- To generate our model by utilizing Python.

The research aims to answer the following questions:

RQ1 How can cyberbullying and online abuse be detected and prevented on Arabic social media platforms.

RQ2 What are the issues and how to solve them for the processing of the Arabic tweets?

RQ3 What is the best performing model for cyberbullying detection?

1.3 Thesis Contribution

The dissertation’s objectives are to fill the research gap for cyberbullying detection in Arabic social media and focusing on the dialect Arabic. It shows and examines various deep learning calcification models for Arabic social media cyberbullying reviews. We also seek to enhance these DL models to gain superior results for the Arabic language. We supply a new dataset for cyberbullying detection for Arabic social media.

1.4 Dissertation Organization

The dissertation is organized as follows:

- **Chapter 1: introduction:** presents an overview of the cyberbullying
- **Chapter 2: Related work:** proposes a summary of the studies that has been deducted in this domain.
- **Chapter 3: Background:** review of cyberbullying and the fundamental concepts engaged. It investigated several cyberbullying technique designs, the importance of cyberbullying detection, and its challenges.
- **Chapter 4: Dataset Corpus:** collect and analyze the datasets that are utilized in this thesis.

- **Chapter 5: Methodology:** illustrates the data collection methodology.
- **Chapter 6: Evaluation and Results:** shows the experimental results that we gained by implementing deep neural network techniques which are CNN, RNN and the combined CNN-RNN and make a comparison between them.
- **Chapter 7: Conclusion And Future Work:** it concludes the dissertation and supply potential orientation for future research.

2. CHAPTER TWO: RELATED WORK

This section presents a literature review of various bullying classifications and sentiment analysis that researchers proposed in their works. The majority of the previous works deducted in the cyberbullying and sentiment analysis can be categorized into three groups: machine learning, deep learning and sentiment orientation. The previous works of cyberbullying and sentiment analysis mostly utilized machine learning techniques, for instance, Naïve Bayes (NB), Decision Tree and Support Vector Machines (SVM). Scholars tended to neural networks for cyberbullying detection by utilizing CNN, LSTM or RNN.

2.1 Arabic Sentiment Analysis Using Machine Learning and Neural Network

Nowadays, various researches have been used in Machine learning and deep learning in the field of NLP, involving sentiment analysis in several languages. Despite all the research performed on English sentiment analysis, few studies have been conducted on Arabic social media.

It is obvious that analyzing such huge data to enhance products, services, and meet customer needs is quite difficult [6]. Thus, employing cutting-edge sentiment analysis tools, the evaluation process must be automated. The majority of earlier efforts concentrated on text-based, mono-modal sentiment analysis. The most recent support vector machine (SVM)-based classification approach is described and validated in this research utilizing a fresh Arabic multimodal dataset. The authors of this study considered SVM as a classifier without considering any other classifiers to validate the results.

The purpose of [7] study is to enhance the sentiment analysis's labeling procedure. There were two strategies used. In order to build a framework of trustworthy Twitter messages with positive, negative, or neutral feelings, a neutral class was first added. The second strategy involved relabeling in order to streamline the labeling procedure. The labeling procedure in this study only applied to seven positive or negative random features: "profits" (ارباح), "losses" (خسائر), "green color" (باللون الاخضر), "growing" (زيادة), "distribution" (توزيع), "decline" (انخفاض), "financial penalty" (غرامة), and "delay" (تأجيل). Twenty tweets out of the 48 that were recorded and evaluated had their labels changed, which decreased the categorization error by 1.34%. The researchers in the previously reported study had limitations and restrictions in the size of the data.

The "OMAM" systems that we created for SemEval-2017 challenge 4 are discussed in [8]. For subtask A, the authors assessed English cutting-edge techniques using tweets in Arabic. With regard to the remaining subtasks, researchers presented a topic-based methodology that takes into account subject specificities by anticipating the topics or areas of incoming tweets and using this knowledge to forecast their emotion. The results show that translating the English state-of-the-art technique into Arabic has produced good outcomes without major improvements. The topic-based approach ranked the first place in subtasks C and E and second in subtask D. the results showed accuracy of 41% which can be improved more in addition to the fact that the researchers implemented single type of classification which is LSTM without considering any comparison between other types of classifications for results evaluations.

investigation of Arabic Twitter characterization in [9] intended to enhance knowledge of the subject while also facilitating a better understanding of Arabic Tweets. Also, studies examined the effectiveness of the two schools of machine learning, namely the feature engineering technique and the deep learning methods, on Arabic Twitter. Researchers took into account models that have attained state-of-the-art performance for English opinion mining. Findings demonstrate the benefits of deep learning-based methods and underline the necessity of employing morphological abstractions to deal with the intricate morphology of Arabic. The results obtained 58% accuracy which can be enhanced more.

Convolutional neural networks (CNNs) for ASA are examined in [10] using a system we developed called CNN-ASAWR on the ASTD and SemEval 2017 datasets. We investigate the impact of different word representations that are unsupervised and acquired from unlabeled corpora. Without utilizing any sort of hand engineering features, experimental findings showed that we were able to surpass the prior state-of-the-art methods on the datasets. The study focused on CNN classifiers without considering other forms of classification to validate the results.

The tweets in [11] study's Arabic-language Twitter corpus from Jordan are labeled as being either positive or negative. It looks at various supervised machine learning sentiment analysis methods as they are implemented to social media posts made by Arabic users on topics of general interest in either Modern Standard Arabic (MSA) or Jordanian dialect. A variety of weight schemes, stemming, and N-grams terms approaches, as well as

other circumstances, are tested in experiments. The experimental findings show the best case for each classifier, and they show that the SVM classifier, which uses the term frequency-inverse document frequency (TF-IDF) weighting scheme and stemming through Bigram's feature, outperforms the Naive Bayesian classifier. Also, the outcomes of this study fared better than those of similar previous studies. The authors in this study were restricted with limited size of data.

In the research [12], the authors provided a set of characteristics that were combined with a sentiment analysis method constructed using machine learning by utilizing Complement Naïve Bayes (CNB), and used on datasets from social media in Egyptian, Levantine, Saudi, and MSA Arabic. An Arabic Emotion Lexicon was used to determine many of the suggested traits. Additionally, the model includes emoticon-based features and input text-related information, such as the number of text segments, the length of the text, whether or not the text finishes with a question mark, etc. among six of the seven datasets the researchers tested on and which are all benchmarked, the authors demonstrate that the offered features have improved accuracy. The researchers of this investigation applied (CNB) solely without considering any other models for evaluation purposes.

In [13], the authors introduced Hotel Arabic-Reviews Dataset (HARD), the largest Arabic book reviews dataset for machine translation and sentiment analysis. 490587 hotel reviews from the Booking.com website are included in HARD. Each entry includes the Arabic-language review text, the reviewer's star rating (from 1 to 10) and other details about the hotel/reviewer. They provided both the entire unbalanced dataset and a portion that is balanced. The researchers built six well-known classifiers utilizing Modern Standard Arabic (MSA) and Dialectal Arabic to analyze the datasets (DA). Researchers evaluated the polarity and rating classifications of the sentiment analyzers. The authors restricted themselves to a limited set of domains.

Despite that, the research work appeared by the unsupervised method is restricted by the data size and the area of the lexicon which restricts the generalizations of the machine method, especially when implementing deep learning methods. Additionally, the works demand knowledge and expertise in this area. Rather than that, the fact of utilizing a sole single classification model without considering a comparison with other models, affects the outcomes results.

2.2 Arabic Cyberbullying Using Machine Learning and Neural Network

Because researchers encountered numerous challenges in detecting cyberbullying in the Arabic context, machine learning techniques are utilized to automatically detect the cyberbullying. Furthermore, it assists the agencies in the government in quickly resolving issues and creating a secure and safe virtual world environment.

In [2], the authors proposed the Naïve Bayes (NB) classifier algorithm for detecting cyberbullying on Arabic platforms. They trained and tested the classifier on a real data set that gathered from Twitter and YouTube. Their Arabic corpus included 25,000 comments that had been manually labeled as either bullying or non-bullying. They implemented the NB classifier for detecting CB, where the obtained accuracy rate was 95.9%. The researchers obtained the results with only a single model without comparing the results with other models.

[14] utilized a feedforward neural network for Arabic cyberbullying detection, where tweets used as the dataset. The researchers altered several parameters in the neural network to make changes and gain better accuracy. The parameters consist of the hidden layers numbers, the epoch's numbers, and the size of the batch. After various training experiments, the researchers have achieved better performance after a few epochs. 16 optimum batch size and the 7 hidden layers were given the best results with 94.56% accuracy. From the results, the authors didn't compare the results with other models for validation.

[15] This strategy is built on data mining techniques that are implemented to a dataset collected from comments on Arabic Facebook postings, and it also presented an algorithm to gauge the cyberbullying severity in a comment. The results obtained an accuracy of 77% in cyberbullying detection by applying Support Vector Machine classifier compared with Adaptive Boosting technique which acquired the highest Precision rate of 94%. The research achieved the objective target but with limitation and restriction in the size of the dataset beside that they focused only on the Syrian Slang without considering any other dialects.

Several classifiers were implemented in this research, SVM model and a radial function kernel, where Lexical features alongside pre-trained embedding were studied in [16]. The results obtained from SVM were the best with 88.6%. The authors studied the SVM sole without comparing results with other classification methods for the evaluation purposes.

With the intention of serving as a benchmark dataset for the automatic detection of Tunisian hazardous contents on social media, [17] offered the Tunisian Hate and Abusive Speech (T-HSAB) dataset. We give a thorough analysis of the data gathering procedures and how we create the annotation standards to ensure accurate dataset annotation. The annotations consistency was later highlighted by a thorough study of the annotations, which used the metrics of Cohen's Kappa (k) and Krippendorff's alpha (α) for annotation agreement. Similarly, [18] presented research aimed at finding offensive words on Arabic social media. With the use of typical patterns seen in harsh and disrespectful interactions, researchers extracted a list of profane terms and hashtags. Also, they categorized Twitter users based on whether or not they utilize any of these words in their tweets. Using this classification, authors increased the list of profane words, and presented the findings using a newly constructed dataset comprising categorized Arabic tweets to (obscene, offensive, and clean). The researchers in the previously mentioned studies, considered the difference in Arabic dialects and its impact on detecting offensive text, which limited their focus to a single dialect solely.

[19] generated a sizable Arabic dataset for obscene and offensive language. In order to understand how abusive language is used by Arabic speakers and what themes, dialects, and gender are most frequently connected with offensive tweets, researchers analyze datasets. To get strong findings ($F1 = 79.7$) on the dataset, they conduct numerous tests utilizing Support Vector Machine algorithms. The authors applied only one classifier which is SVM, where the results are not compared with other classifiers to validate the proposed method.

For sub-task A of the Multilingual Offensive Language Identification shared task (OffensEval 2020), a component of the SemEval 2020, we discuss in this study how we plan to use pre-trained BERT models with convolutional neural networks. We demonstrate that CNN and BERT work better together than alone, and we underline the value of employing pre-trained language models for downstream tasks. Our system placed fourth in Arabic with a macro averaged F1-Score of 0.897, fourth in Greek with a score of 0.843, and third in Turkish with a score of 0.814. We also share with the community ArabicBERT, a set of pre-trained transformer language models for Arabic.

For sub-task A of the Multilingual Offensive Language Identification shared task (OffensEval 2020), a component of the SemEval 2020, [20] discussed in this study how to

use pre-trained BERT models with convolutional neural networks. Scholars demonstrated that CNN and BERT work better together than alone, and they underlined the value of employing pre-trained language models for downstream tasks. The proposed system placed fourth in Arabic with a macro averaged F1-Score of 0.897, fourth in Greek with a score of 0.843, and third in Turkish with a score of 0.814. They also shared with the community ArabicBERT, a set of pre-trained transformer language models for Arabic. The authors restricted and limited the proposed study with the dataset size.

2.3 English Cyberbullying Using Machine Learning and Neural Network

The research paper [4] implemented machine learning using SVM and (NB) classifiers to detect and prevent bullied comments and tweets. The data was collected from Twitter and they gained 71.25% as a highest accuracy which can be enhanced more.

Several classifiers were implemented in this research, SVM model and a radial function kernel, where Lexical features alongside pre-trained embedding were studied in [16]. The results obtained from SVM were the best with 88.6%, which can be improved more.

A hierarchical attention network that considers these elements of cyberbullying in order to detect it was suggested in [21]. The main distinguishing features of our method include I a hierarchical structure that mimics the social media session structure.; (ii) stages of algorithms used at the phrase and post level, allowing the model to pay varied levels of focus to words and posts, based on the scenario; and (iii) a cyberbullying recognition mission that additionally predicts the duration between two adjacent remarks. The suggested design beats the state-of-the-art approach, according to tests on a dataset collected from Instagram, the social networking site where the greatest number of individuals have suffered extreme cyberbullying. The authors of the previously mentioned studies didn't determine which kind of cyberbullying they are identifying.

[22] demonstrated how DNN models may be utilized for cyberbullying detection on a variety of topics across numerous social media platforms (SMPs). For all three datasets, these models outperformed the state-of-the-art results when combined with transfer learning. Authors conducted the experiments with a limited size of data.

The ability to categorize a tweet as racist, sexist, or neither is how [23] described this assignment. This assignment is extremely difficult due to the intricacy of the natural language constructs. To manage this complexity, the authors conducted extensive experiments with a variety of deep learning systems. The tests on a dataset consisting of 16K labeled tweets demonstrate that these deep learning techniques perform about 18 F1 points better than the most advanced char/word n-gram techniques. The obtained high F1 scores are because of the over-fitting [24].

In order to identify whether or not the text contains hate speech, [25] suggest employing a deep learning technique with the Recurrent Neural Network (RNN). There are 652 records classed as hatred in the 1235-record total twitter dataset that employs it. The authors focused on a single type of classification which is (RNN) to detect the hate speech without considering comparing other classification techniques which lower the evaluation of the proposed method.

In [26] research aims to investigate how Natural Language Processing can be used to identify hate speech. Also, this research uses a dataset to apply a recent method in this area. Convolutional Neural Networks (CNN) are a type of deep learning algorithm that have been proposed because they outperform traditional methods for text classification issues. Each tweet is categorized by this classifier into one of the three categories in the Twitter dataset: hate, offensive language, or neither. This model's effectiveness has been evaluated by measuring its accuracy as well as its precision, recall, and F-score. The final model has a 91% accuracy, 91% precision, 90% recall, and a 90% F-measure. The authors concentrated their study on CNN for the purpose of hate speech recognition without taking into account the comparison with other techniques which would decrease the evaluation of the presented model.

In conclusion, after reviewing the previously mentioned approach, we can clarify that our research will resolve the majority of the restrictions of the previous and recent researches in the cyberbullying detection of Arabic social media.

3. CHAPTER THREE: LITERATURE REVIEW

3.1 Background

In the recent days, the global has seen various revolutions that were made doable by Social Media. It is a highly effective innovation of our lives, and is an impressive way to increase the borders of the person's experiences and become active on the social. Though, social media is a two-sided arm. Several anti-social attitudes, in other words bullying, are spotted on social media. Bullying has recently been part of rising up. Furthermore, this is not restricted to children or youth; anyone can be a bullying victim.

3.1.1 Cyberbullying Definition

Cyberbullying is officially definite as “willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices” [27]. Another definition is “an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself” [28]. In brief, bullies normally avail the use of technical communication for harming people. This hurt may be inspired by annoyance, prevention, payback, or from a basic wish to manipulate others or to sense more power [29]. Occasionally, children cyberbully others to fit with their own low self-confidence and/or to cope with their analogues [29]. There are many types of bullying. When any form of bullying is published, it is very hard to retrieve these posts back off from the social media websites. This can happen any time of the day, even day or night time for the whole week. This can affect the victims whenever they are, by themselves, in public areas, at school, or even in the sports yards [27]. Cyberbullying authorizes a bully to embarrass and offend the victim in online societies without ever being predictable. Moreover, punishment fear or being a social castaway prevents victims and bystanders from reporting the incidents. This becomes a tough problem to manage. Cyberbullying attitude is not only inadmissible, but can also lead to tragic consequences. The researchers in [30] show that “critical impacts occurred in almost all of the respondents’ cases in the form of lower self-esteem, loneliness and disillusionment and distrust of people: The more extreme impacts were self-harm and increased aggression towards friends and family”. They also mentioned that many victims progressed “coping strategies.” Several times, victims attempted to cope with cyberbullying all by themselves, which led to a tense case. Moreover, it is difficult for victims’ parents to realize what is going on with their kid

online. For support systems to be able to assist the victim, they have to recognize the cyberbullying or any indicator of it at the initial. They should not anticipate the victim to inform them regarding cyberbullying. This demands for automated cyberbullying detector programs that could warn family's members about cyberbullying.

3.1.2 Cyberbullying on Social Media

Compared to spam detection mails, where several recipients receive the same broadcasted message, detecting cyberbullying, which is more individualized and context directed, is much harder. Most cyberbullying has been determined to concentrate on specific subjects consisting; racism and ethnicity, sexual identity and sexuality, shape of the body and look, intellect, and sociable incorporation and rejection. Recognizing whether a textual expression and its sentiment can cope with these topics, and the tone is positive or negative, would be essential in detecting probable cyberbullying comments [31] [32].

3.1.3 Cyberbullying Forms

This part specially concentrates on the features of cyberbullying detection in social media. It comprises user personality's features, sentiment features, emotional features and Twitter-founded features.

There are various forms of cyberbullying reported in the social media as following [33]:

- 1- **Exclusion**: exclude or eliminate somebody from a group in social media. This can cause a victim feeling dejected and secluded.
- 2- **Harassment**: sending frequent harmful and insulting online messages to a victim where messages may include threats.
- 3- **Cyberstalking**: sending fake accusations and threats to the victim about him and his loved ones which frightening him.
- 4- **Outing**: posting private and personal information about the victims in the public on social media without their permission.
- 5- **Frapping**: when someone uses your account on social media and claims they are the account's owner to post improper materials. In this case, the victim is strapped to online content that can destroy his reputation.

6- **Trolling**: posting contentious comments on social media purposely to upset others on social media to hurt others.

7- **Dissing**: Posting private information or photos about another person with the intent to damage their reputation or fame.

8- **Flaming**: Initiating an online fight with the victim.

9- **Denigration**: Publishing false gossip about someone else.

10- **Trickery**: tricking people to the trust point so they can give out secrets, thus getting humiliated or getting made fun of.

11- **Masquerade**: the bully pretends to be someone who is not. Cyberbullies can set up fake online profiles on behalf of victims. They can use these profiles to publish false content in their victims' names without the victims' consent.

12- **Catfishing**: stealing somebody's identity online and creating a false profile to fool others. The majority of the studied literature does not identify which form of cyberbullying they are detecting in their studies. Nonetheless, the most public form of cyberbullying is online harassment as in the literature [34] [35] [21]. There are sub-forms of harassment declared in the revised literature such as Aggression [36], [37] and Toxicity [38]. Figure 1 illustrates the cyberbullying forms.

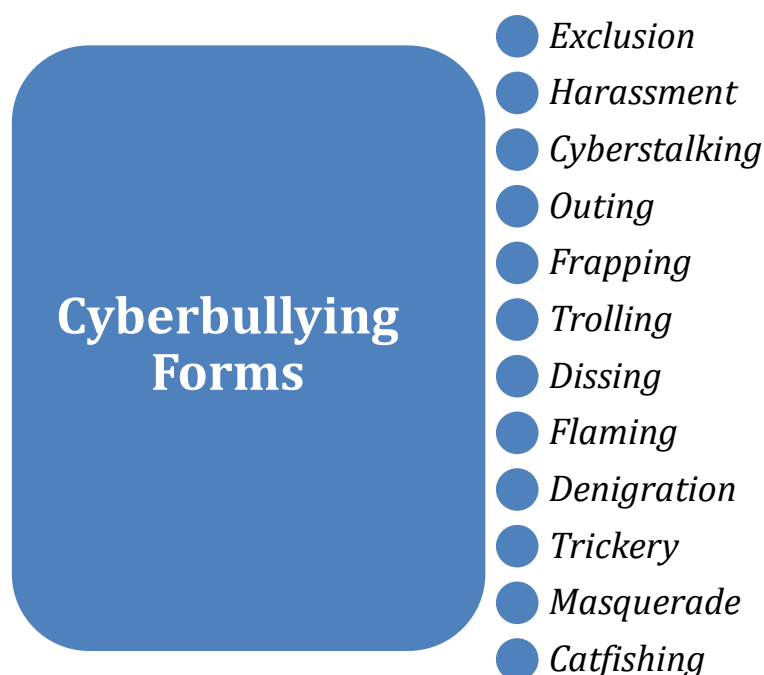


Figure 1: Cyberbullying Forms

3.1.4 Sentiment Analysis

Sentiment analysis, also known as opinion mining, is the use of natural language processing, text analysis, and computational linguistics to identify and extract subjective information from source materials. It is used to determine the attitudes, opinions, and emotions of a speaker or writer with respect to some topic or the overall contextual polarity of a document. Sentiment analysis can be used to monitor public opinion on social media platforms, analyze customer feedback, and even assist in political campaigns. It is an important tool for businesses and organizations to understand how their products and services are perceived by the public.

Sentiments are represented by the ideologies or concepts excited by the sensations fastened with something, usually classified as positive, neutral or negative [39]. It is obtained from the demonstrated opinion in a document, sentiments or text subjectivity on social media. The method in which the unformed data are computationally treated, is addressed as sentiment analysis, and can be classified into machine learning method, lexicon-based technique and hybrid approach [40]. Machine learning method implements algorithms such as SVM, Naïve Bayes, Decision Tree, etc., while the lexicon-based technique is based on sentiment lexicons (i.e. dictionary of opinion expression and phrases with the designated polarities and severity) for measuring a text sentiment. In the situations of cyberbullying, sentiment has been utilized to differentiate between no bullies, victims and bullies [9]–[11]. For illustration, [41] specified victims of cyberbullying utilizing their sentiment scores as victims frequently face negative feelings or emotions such as loneliness, anxiety and depression.

3.1.5 Emotion

Contrasting sentiment analysis, emotion analysis detects forms of sensations via the text expression, such as outrage, disgust, panic, happiness, sorrow, and surprise. Three popular techniques occur in textual emotion discovering, known as, keyword-based (i.e. utilizing dictionaries' synonyms and antonyms type), learning-based (based on the classifiers that earlier trained) and hybrid (mixture of keyword and learning techniques) [44]. Emotion analysis have been implemented in several fields, for instance, novels [45] as well as suicide notes [46], [47]. Authors in [41] determined seven emotions, which are popular in cyberbullying attitudes, and those are empathy, embarrassment, anger, relief, pride, fear, and sadness. To be particular, the researchers discovered accusers give more fright but fewer

anger, comparing victims beside reporters who appeared to face more grief and comfort. [48] checked the existence of anger, fright and sadness in cyberbullying, though, no magnificent effect was recorded even though a development on the accuracy in specifying aggressive attitude.

3.1.6 Arabic Language

Arabic is a language that is spoken by over 300 million people in the Middle East, North Africa, and the Horn of Africa. It is the official language of many countries, including Saudi Arabia, Qatar, Kuwait, Iraq, Egypt, Lebanon, and the UAE. Arabic is a Semitic language, related to Hebrew and Aramaic. It is written in a form of the old Aramaic alphabet, which was developed in the Middle East over two thousand years ago. Arabic is written from right to left and consists of 28 alphabets, and its syntax is quite different from English. Arabic is also an important language in the Muslim world, as it is the language of the Qur'an, the holy book of Islam. Arabic is also used in many Muslim countries as a second language.

The complicated morphology of Arabic linguistics excludes researchers from studying it. There are three types of Arabic language: The first one is the Classical Arabic which is the ancient language and is the language of the Holy Quran, and Muslims utilize it in their prayers. The second type is the Modern Standard Arabic (MSA) that is considered the official language spoken in news broadcasts, formal documents, and education, and is well known by the Arabs. The last type of the Arabic Language is the Dialect accents and are the spoken linguistic among the Arabic people. The dialect accents differ from place to another where the Arabic word consists of around 10 dialects based on their location [49] [50]. The difference in the dialects created differences in meaning of some words, for instance the "Yetqalash" which is expressed in Yemini for a compliment while in Morocco it means an offensive. Due to the Arabic language nature, it is considered challenging, where there is no Capitalization in its words [49]. Arabic morphology causes a lot of confusion where the Arabic corpus is considered rare. Arabic linguistics is ranked the 7th over the universe, and it is growing widely over the social media which motivate the research in the Arabic linguistic domain.

A comprehensive search was conducted on convenient publications and articles and rare previous publications was performed for cyberbullying detection using Arabic language textual except few papers in the area of machine learning and Natural Language Processing (NLP). The earlier research accomplished in Arabic applied text preprocessing or

classification of the text. In the [51], the researcher presented a key phrase extraction technique to join multiple key phrase extraction techniques with ML methods. The outcomes from the key phrase extraction techniques are utilized as features to the ML techniques while the ML method classifies these features as even a key phrase or not. They contrast their results by utilizing three ML algorithms: Linear Discriminant Analysis and SVM [52], and Linear Logistic Regression [53]. Their results proved that SVM presents the optimum results in key phrase extraction amongst the others. Much research had been performed in Arabic known entity extraction, for example Named Entity Recognition for Arabic (NERA) [54], to clarify the appropriate names in Arabic text and documents. NERA utilizes a list of named entities and corpus compiled from a variety of resources; the evaluation was done via recall, precision and F1. The outputs showed satisfaction results; 86.3%, 89.2% and 87.7% for person named entities. Spam emails filtration using Arabic and English language was performed in [55]. They performed the proposed technique on pure Arabic, pure English and a combination of Arabic and English emails. Many ML methods were utilized, consisting K-Nearest Neighbor (K-NN), NB, SVM [56] and Neural Network. The performance of the proposed technique was evaluated across all the previously mentioned methods and SVM showed the premium results in the pure English among the others. The system performance in the Arabic emails was not satisfactory compared to the English, because of the complicated nature of the Arabic language. The research also confirmed that stemming Arabic wording improved the efficiency of the classifiers, while NB proved the best performance with 96.78% Recall and with F1 of 92.42% measuring. In addition, there are some efforts for detecting spam emails in social media, for instance twitter. Researchers in [57] enhanced a system to detect spam messages in Arabic tweets. The proposed system performed remarkable results based on the accuracy, precision and recall measurements.

Sentiment analysis is considered one of the text classification groups. Sentiment Analysis sorts the textual into positive, negative or natural [58]. Sentiment analysis was founded by [59] using Facebook comments in Arabic. They constructed a corpus from 6000 comments extracted from Facebook, preprocessing the data, then implemented classifications to specify the sentiment type. Three types of classifiers were implemented: NB, Decision Trees and SVM. The results proved the effectiveness of SVM compared with the other methods with accuracy 73.4 %. Another research was conducted for sentiment analysis for Arabic tweets done by [60]. Their specific contribution was done on dialects and Arabizi. They combined K-NN, SVM and NB to classify the comments, and that NB

provided the best results. In [61], the researchers utilized the dialectic Arabic text for detecting the sentiment. For the detection, they implemented two methods: translating the dialectical terms into MSA, then the detection is performed by MSA lexicon. The second way was via dialectical lexicon detection. SVM besides NB were used as classifiers to detect the polarities either positive or negative. The obtained results proposed enhancement in Precision, Recall and F-measure for translating into MSA over the other method. In [58], the sentiment analysis was performed on Arabizi as well. The proposed system firstly converted the Arabizi text into Arabic via utilizing their technique. They used the crowdsourcing tool to label the data then implement SVM and NB as classifiers. A comparison was conducted between SVM and NB, and the results proved the outperformed SVM. In Saudi Arabia, the Arabic tweets were analyzed by [62]. They utilized a hybrid method for analyzing tweets. The hybrid technique consisted of constructing a classifier besides training it via a one-word dictionary. The results were compared with the supervised learning method and the results proved the superiority of the hybrid technique. An impressive research was conducted on stemming for Arabic linguistics. Stemming is a process of decreasing the word to its root [63]. Various Arabic stemmers are obtainable, consisting rule-based stemmers, for example Khoja [64] besides light stemmers where Light stemmers eliminate letters from words – suffixes and affixes- without knowing the words’ roots.

3.1.7 Arabic Text Cyberbullying Types

Cyberbullying is a phenomenon that is becoming harsher worldwide, and regrettably, it is also a concern in Arabic-speaking nations. The Arabic language is expressed as a delicate and complicated language. Arabic slang tweets were labeled by appending “bullying attitude” which classified based on the following cyberbullying forms [15]:

- 1) **Sexual Cyberbullying**: this form contains the embarrassing words to the persons with sexual offense for instance as “Slut” "فاسقة".
- 2) **Physical Sexual Cyberbullying**: this type of cyberbullying consists of particularly the body's private sexual portions.
- 3) **Religious cyberbullying**: this form consists of the expressions that aim at the religious and doctrine of the victims, such as: "خاين" which means “traitor”.
- 4) **Political Cyberbullying**: this type includes the bullying expressions that target the behavior in politics, for instance: "ذنب الحكومة" which means “the government tail”.

- 5) **Looking cyberbullying**: this type represents the words that depicts the victims looking, such as "أنت بشعة" which means "you are ugly".
- 6) **Animal Phrases Cyberbullying**: this type is related to the description of the people with the type of animals, for instance "بومة" which means "owl".
- 7) **Racism Cyberbullying**: this cyberbullying is pointed to the racialism or ethnic nationality of the persons, for example "عبد أسود" which means "nigger".
- 8) **Sectarian Cyberbullying**: the phrases of this cyberbullying arise between two groups who have political or cultural conflict, such as "شيعي" which means "Hateful Shiite".
- 9) **Cross-Cultural Cyberbullying**: this type of cyberbullying deals with the abusers' endeavor to offend the victim's culture [67], for example "شحاد" which means "Beggar".
- 10) **Psychological cyberbullying**: is the bullying associated with psychology, for example "انت واحد ما بتستحي" means "shame on you".
- 11) **Praying Negatively for Person**: for instance "الله ياخذ روحك" which means "god take your soul".
- 12) **General Cyberbullying**: this cyberbullying does not belong to the previously mentioned forms, such as "تفه عليك" which means "spit on you".
- 13) **Non cyberbullying**: the text does not consist of any expressions associated with cyberbullying/ attitude, for example "انا بعرفك" which means "I know you".

Table 1 represents the cyberbullying types and examples of each one. This categorization of cyberbullying attitude has been validated by Arabic language experts and Arabic native speakers.

Table 1: Cyberbullying Types and examples

Cyberbullying Types	Examples	
	Arabic	English
Sexual Cyberbullying	فاسقة	Slut
Physical Sexual Cyberbullying		
Religious cyberbullying	خاين	Traitor
Political Cyberbullying	ذنب الحكومة	Government Tail
Looking Cyberbullying	أنتِ بشعة	You are Ugly
Animal Phrases Cyberbullying	بومة	Owl
Racism Cyberbullying	عبد أسود	Nigger
Sectarian Cyberbullying	شيعي حاقد	Hateful Shiite
Cross-Cultural Cyberbullying	شحاذ	Beggar
Psychological Cyberbullying	انت واحد ما بتستحي, غبي	stupid, Shame on you
Praying Negatively for Person	الله ياخذ روحك	God take your soul
General Cyberbullying	تفه عليك	Spit on you
Non Cyberbullying	انا بعرفك	I know you

3.2 Deep Learning Approaches

3.2.1 Convolutional neural networks

Convolutional neural networks (CNNs) are one of the effective deep learning techniques used for computer vision and image recognition applications. CNNs are built to automatically recognize patterns and characteristics inside images and are inspired by the way the brain interprets visual information.

CNN uses many layers to analyze the data, doing away with the necessity for human extraction. CNNs are a potent NLP technique that produced excellent results. Tasks requiring computer vision frequently employ CNNs. It may be used to several pictures, documents, audio files, and films. The CNN is frequently used in hierarchical data categorization to identify patterns.

CNNs are indeed a mathematical expression that stimulates the neural networks in the brain. Others can utilize three straightforward levels, such as input, output, and hidden layer, while some just use the two main layers—input and output layers. 2019 (Ibrahim et

al.). The previous several years have seen a lot of NLP and sentiment analysis research, with promising outcomes. The CNN and LSTM algorithms clearly beat the single model. (SVM, NB and ME). In order to represent the grammar as a collection of words, CNN employed by utilizing a single layer of convolution.

CNN Convolutional Layer

The convolutional layer, which applies a group of teachable filters or kernels to the input picture, is the fundamental component of a CNN. In order to create a feature map, each filter performs a dot product operation as it moves over the input picture. The feature map displays where on the input picture that filter was activated.

In order to create a 3x3 feature map, a 3x3 filter is applied to a 5x5 input picture in figure . To determine the value of each pixel in the feature map, the filter traverses through the input picture, executing a dot product at each place.

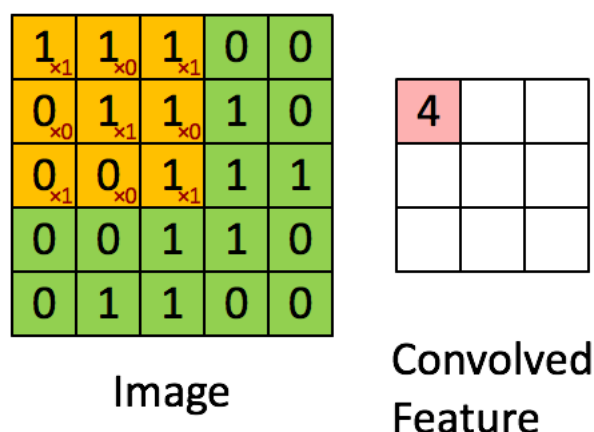


Figure 2: CNN Convolutional Layer

CNN Pooling Layer

The rectified linear unit (ReLU), which brings non-linearity into the network, is one non-linear activation function that is then applied to the output of the convolutional layer. A pooling layer follows, which lowers the feature maps' dimensionality and increases the network's computational efficiency.

The graphic in figure 3 illustrates how a 4x4 input feature map is subjected to a 2x2 max pooling operation to produce an output feature map. The spatial dimensions of the

output are decreased by the pooling procedure, which chooses the largest value in each 2x2 window of the input feature map.

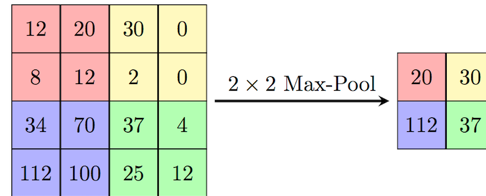


Figure 3: CNN Pooling Layer

Following that, further convolutional and pooling layers are applied to the pooled feature maps, each of which learns increasingly intricate patterns and information from the layer before it.

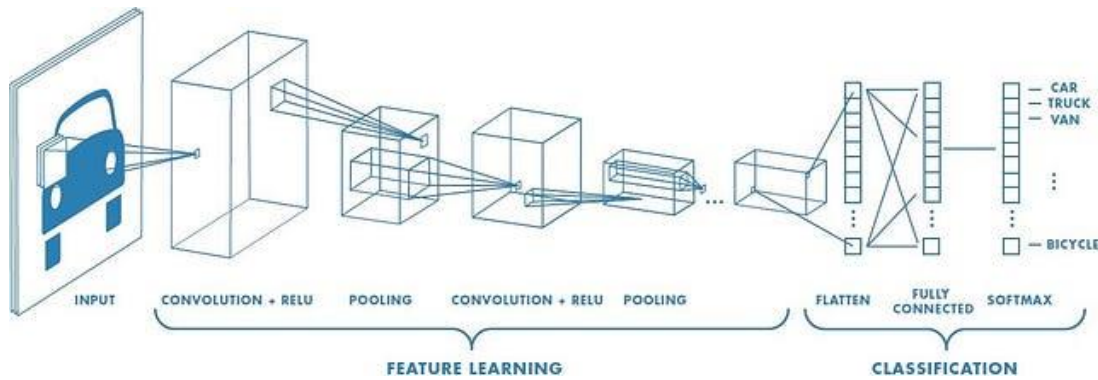


Figure 4: CNN Architecture

Convolutional Neural Network Architecture:

One or more fully connected layers then execute classification or regression tasks based on the learnt features after flattening the output of the final convolutional layer into a 1D vector. The architecture of a typical CNN is depicted in figure picture, which includes a number of convolutional and pooling layers followed by one or more fully connected layers for classification or regression.

Visualizing Feature Maps:

The picture in demonstrates how to view the feature maps that a CNN has learnt in order to recognize the patterns and features that the network has discovered. In the first convolutional layer, each row represents a separate filter, and the columns display the activations of that filter across various input pictures.

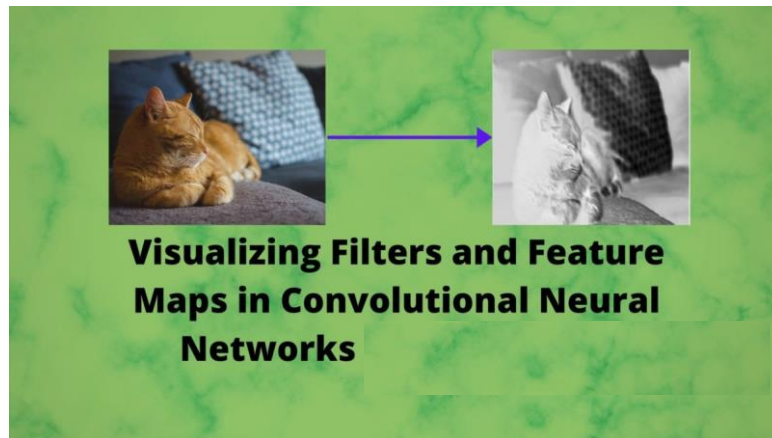


Figure 5: Visualizing in CNN

3.2.2 Recurrent Neural Networks (RNN)

A particular class of neural network known as recurrent neural networks (RNNs) is designed specifically for processing sequential data, including time series, audio signals, and spoken language. RNNs are intended to keep a memory of prior inputs and utilize that knowledge to forecast what will be inputted in the future, unlike feedforward neural networks that analyze each input individually.

A Recurrent Neural Network's (RNN) key characteristic is the presence of at least one feed-back link, which allows activations to cycle around in a loop. This gives the networks the ability to process time and learn sequences, for example, to recognize and reproduce sequences or to anticipate temporal relationships between events. The architectures of recurrent neural networks can take many distinct shapes. One typical version has an additional loop and a conventional Multi-Layer Perceptron (MLP). They have some type of memory and can make use of the MLP's potent non-linear mapping capabilities. Others may have more uniform architectures with every neuron connecting to every other neuron and/or stochastically activated neurons. Similar gradient descent techniques to those

leading to the back-propagation method for feed-forward networks may be used to train for basic structures and deterministic activation functions.

In sequential tasks like processing voice and natural language, the current input data is constantly dependent on the applied inputs from the past. Finding the connection between the current input and the previously applied inputs is the job of RNNs. RNNs can theoretically employ information sequences of limitless length, but in reality they can only go back a short distance.

In the illustration above, an RNN is unfurled into a complete network. Just repeating the same network layer structure throughout the whole sequence is what we mean when we say we are unfolding.

The input at time step t is X_t . The vector X_t can be any size N . At time step t , the concealed state is A . It is the network's "memory." Based on the input at the current step and the prior concealed state, it is computed.

$$\text{As shown by } A_t = f(W X_t + U A_{t-1})$$

Here, the input and prior state value input weights are denoted by W and U . Moreover, f is the nonlinearity that is applied to the sum to produce the final cell state.

One of the allures of RNNs is the potential for them to make connections between prior knowledge and the work at hand, for example, by using prior video frames to help interpret the present frame. RNNs would be very helpful if they could accomplish this. Sometimes, all we need to do to complete the work at hand is to glance at the most current data. Consider a language model that uses the words before it to attempt to predict the word that will come next. There is no need for more context if we are attempting to determine the final word in "the clouds are in the sky" because it is very clear that sky will come after it. RNNs can learn to utilize the prior information in such circumstances, when there is a close proximity between the pertinent information and the location where it is required. Theoretically, RNNs can handle such "long-term dependencies" with ease. To solve these kind of toy issues, a person might carefully choose the settings for them. Regrettably, RNNs don't seem to be able to learn them in practice.

3.2.3 Long Short Term Memory Networks (LSTM)

They are a particular sort of RNN, commonly known as LSTMs, and are able to learn long-term dependencies. They are currently frequently utilized and perform incredibly well when applied to a wide range of issues. Intentionally, LSTMs are created to prevent the long-term reliance issue. They effectively function in a way that makes long-term memory retention their default mode. All recurrent neural networks have the shape of a series of neural network modules that repeat. This recurring module in typical RNNs will be made up of just one tanh layer, for example. The weight matrix associated with the connections between the neurons of the recurrent hidden layer can end up multiplying the gradient signal by a significant number of times (as many as the number of time steps) during the gradient back-propagation phase in all traditional recurrent neural networks. This indicates that the learning process may be significantly impacted by the size of the weights in the transition matrix.

The gradient signal may become so small in the case of stochastic gradient descent, when learning either becomes extremely slow or ceases to function, if the weights in this matrix are small (or, more precisely, if the leading eigenvalue of the weight matrix is lower than 1.0). Learning long-term dependencies in the data may become more challenging as a result. On the other hand, if the weights in this matrix are enormous (or, to put it more precisely, if the leading eigenvalue of the weight matrix is larger than 1.0), it may result in a scenario where the gradient signal is so significant that learning may diverge. Exploding gradients is a common name for this.

These difficulties serve as the primary driving force for the LSTM model, which adds a brand-new component called a memory cell. An input gate, a neuron with a self-recurrent connection (a link to itself), a forget gate, and an output gate make up the four essential components of a memory cell. With a weight of 1, the self-recurrent connection guarantees that, without external interference, the state of a memory cell can remain constant from one-

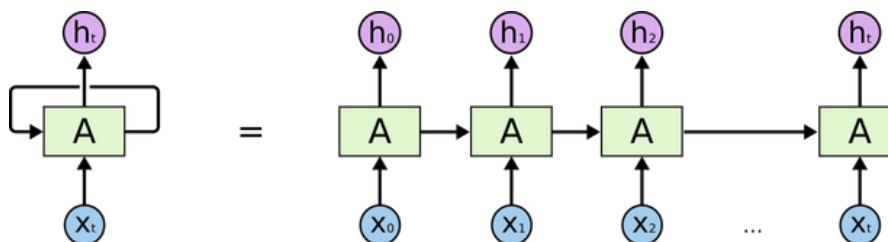


Figure 7: Basic architecture of Recurrent Neural Networks

time step to the next. The gates are used to control how the memory cell interacts with its

surroundings. An incoming signal can be blocked by the input gate or allowed to change the state of the memory cell. The output gate, on the other hand, can either permit or disallow the state of the memory cell from having an impact on other neurons. The memory cell may recall or forget its prior state as needed by modulating the self-recurrent connection of the memory cell through the forget gate as shown in figure 8.

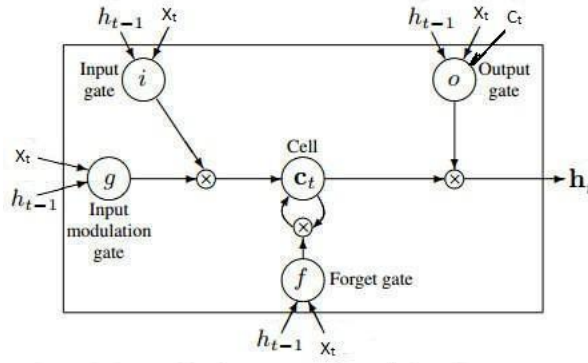


Figure 8: LSTM Single Cell

The input gate i , forget gate f , and output gate o are the three gates in an LSTM. g is the input that has to be updated in the cell at this time step.

According to the recurrent nets concept, each of these gates and the update signal depend on the cell's current input as well as the prior hidden state.

3.2.4 Hybrid CNN-RNN

Convolutional and recurrent neural networks are combined in CNN-RNN, a hybrid neural network design, to analyze sequential data such as time series and text written in natural language. In many different applications, including speech recognition, picture captioning, and video analysis, the CNN-RNN architecture has been extensively used because it can detect both local and global patterns in sequential data.

The CNN-RNN architecture is made up of two parts: a CNN component and an RNN component. Local patterns from the input sequence are extracted by the CNN component, while the long-term relationships between the patterns are captured by the RNN component.

The CNN component is used to apply a series of convolutional filters to the input sequence in order to extract local patterns. Each feature map in the output of the CNN component captures a distinct local pattern from the input sequence.

The RNN component analyses the series of feature maps and captures the long-term relationships between the local patterns after receiving the feature maps. A basic RNN, a gated RNN such an LSTM or GRU, or a bidirectional RNN can all be used as the RNN component. An example of a more sophisticated CNN-RNN architecture is shown in figure 10. In this illustration, a group of photos serves as the input sequence, and the CNN component extracts regional patterns from each image. A series of feature maps, each of which reflects a distinct local pattern in the picture, is the output of the CNN component.

CNNs have been applied to text classification jobs as well as picture recognition problems. Sliding window techniques are used by CNNs to extract local characteristics from the input, enabling them to recognize significant patterns and features at various sizes. Nevertheless, CNNs are less successful at tasks that call for comprehension of the context

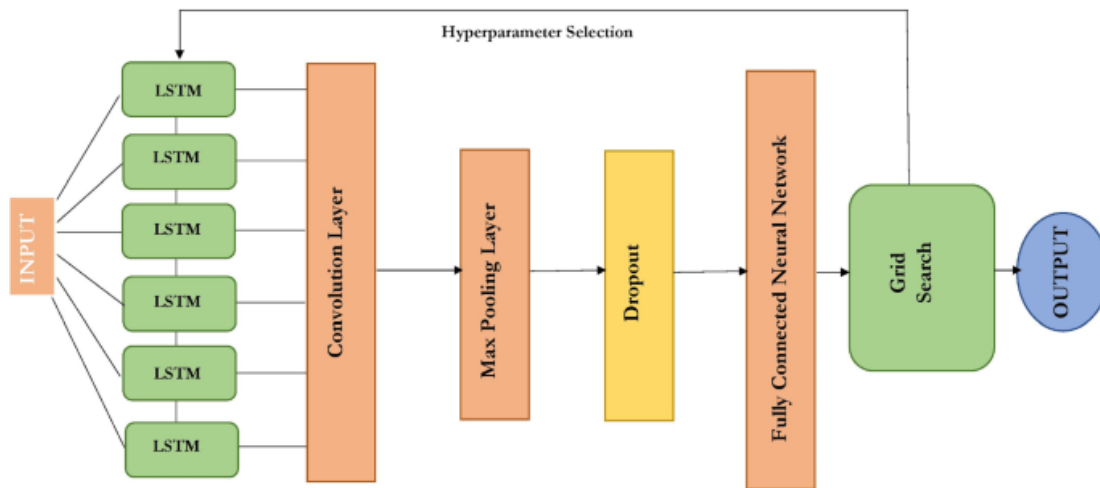


Figure 9:CNN-LSTM Architecture

of the full text because they do not capture long-range relationships between words.

3.2.5 A comparison between deep learning models

In the category of machine learning algorithms known as "deep learning," complex patterns and relationships are learned from big datasets. We'll contrast a few of the most well-liked deep learning algorithms in this article and point out their advantages and disadvantages.

- 1- Convolutional Neural Networks (CNNs): CNNs are typically employed in jobs involving image recognition and video analysis. Convolutional layers are used by CNNs to learn spatial patterns and extract characteristics from pictures. CNNs

can manage changes in object size, shape, and orientation and are noise-resistant. To learn complicated characteristics, CNNs may need a lot of data, and the training procedure might be computationally expensive.

- 2- Recurrent neural networks (RNNs): RNNs are applied to sequential data in time series analysis, speech recognition, and natural language processing. RNNs contain a feedback mechanism that enables them to process each input in the sequence using their internal state. RNNs can handle variable-length sequences and are effective at capturing long-term relationships in the data. Yet, the vanishing gradient problem that RNNs experience might make it challenging to understand long-term dependencies.
- 3- Long Short-Term Memory Networks (LSTMs): LSTMs are a particular RNN type that can solve the vanishing gradient issue. In LSTMs, information is controlled by gating mechanisms while memory cells are used to retain data for extended periods of time. With its ability to handle variable-length sequences, LSTMs excel in modeling sequences with long-term dependencies. Nevertheless, learning complicated patterns with LSTMs can be computationally costly and necessitate a lot of data.
- 4- GANs (Generative Adversarial Networks) are utilized for generative tasks including text creation and picture synthesis. Two neural networks make up GANs: a generator network that generates new data and a discriminator network that aims to separate created data from actual data. GANs are capable of producing high-quality samples and can pick up on intricate distributions. GANs, however, can be challenging to train and may experience mode collapse, in which the generator only generates a narrow range of samples.
- 5- Transformers: The neural network design known as a transformer is generally employed for jobs involving natural language processing. Transformers recognize long-range relationships between words in a phrase via self-attention techniques. Modern performance on a variety of natural language processing tasks, including sentiment analysis, question answering, and machine translation, has been attained using Transformers. Transformers, however, may be computationally costly and need a substantial quantity of training data.

3.3 Data Pre-Processing

In both NLP and SA, the pre-processing stage is important. It assures higher performance in sentiment analysis and helps to enhance the state of the dataset that was gathered. The standard pre-processing techniques include text cleaning, normalization, tokenization, stop word removal, and stemming. The preprocessing procedure can be done in a number of steps, depending on the study purpose and the language used. The effectiveness of Arabic sentiment categorization employing a unique pre-processing step was assessed in researches. Using a customized stop list and employing emoticons are part of the procedures. The results show that by replacing emoticons with customized stop lists and content phrases, Arabic attitude classification may be made more efficient.

A crucial step in enhancing data quality is data cleansing. The task of detecting data polarity is improved. Following data collection, the cleaning stage eliminates misspellings and slang terms. Given the variety of dialects and Arabic language varieties, it is difficult to use this in Arabic text. When it comes to the linguistic domain for NLP, Arabic is a difficult language to work with. It requires sophisticated pre-processing because to its morphological challenges and range of dialects. Delete usernames, hash tags, URLs, punctuation, and extra whitespace as part of a beneficial social media corpus cleaning procedure. Others may eliminate lengthy words, special characters, numerals, and English text depending on the assignment. It is very important to pre-process and assess Arabic unpolished contents.

3.3.1 Data Cleaning

In this part, several data preprocessing steps and cleaning were conducted on the collected tweets or data, to reduce the Arabic language complexity as well as to improve the results quality. The data preprocessing stage was applied by Python. Fig. 11 and the following steps represent the preprocessing and cleaning data structure:

1) Data Collecting

2) Data Uploading and Reading: the collected dataset was loaded to perform the preprocessing steps using the IO standard Python library and was read using a data frame.

3) Noise Elimination:

- Eliminating all non-Arabic tweets and RT (Re-Tweet).
- Remove Arabic and English numbers like (1, 2) and punctuations like (, ?).
- Delete URLs, mention user by @, emotional stickers, hyperlinks and hashtags.
- Removing stop words like (ما , عشان) and Latin characteristics.

- Eliminating repeated letters like (حلوووو) and repeated words () .
- Remove Diacritics and Tatweel.

4) Normalization:

- Substituting the character (ة) with (o).
- Replacing the character (ى) with (ي).
- Substituting the character (آ ,أ ,إ) with (I).

5) Stop words removal: is the method of eliminating words from sentences that only serve to structure language and add little to the primary content. These words include "a" and "are," "the" and "was," for instance. The work of sentiment analysis or categorization is not made easier by these terms, which are typically prevalent. Due to the fact that no essential information is lost, this aids in decreasing cuprous size.

5) Tokenization: separates the raw text string using whitespaces into a list of distinct words, is a crucial step for the majority of NLP jobs. A sentence or document is divided into tokens, which are words or phrases. In languages like English, Arabic, and French, where spaces are used to separate words, it is a straightforward process. In languages like Japanese, Chinese, and Thai, where words are not separated, it is more difficult. Arabic is a language where tokenization may be applied since most words are separated by spaces.

6) Delete instances that are irrelevant: Look for instances that are not pertinent to the current work and eliminate them. Remove instances that don't communicate emotion, for instance, if the aim is to classify sentiment.

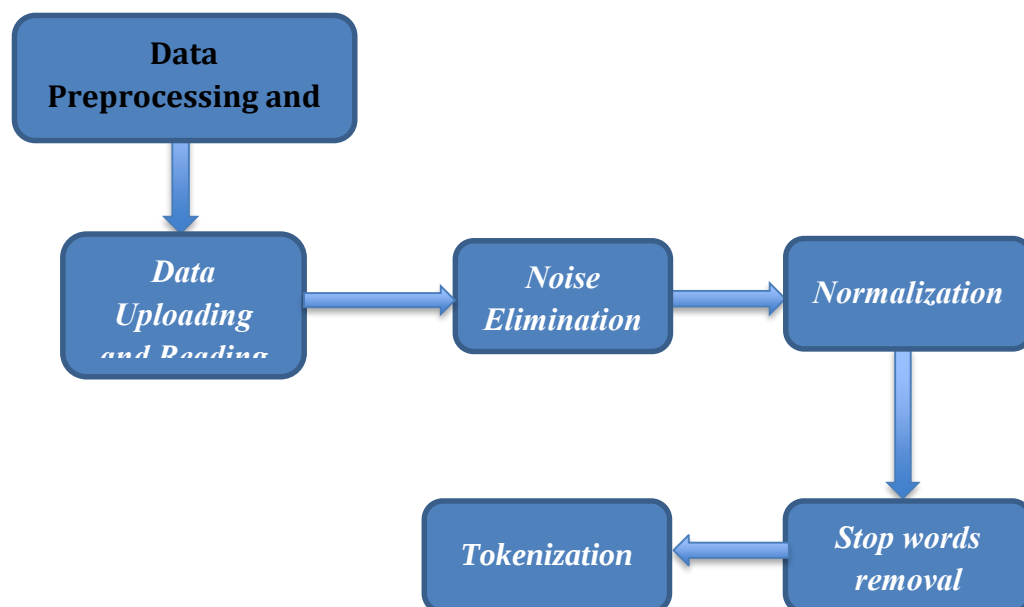


Figure 10:Data Preprocessing and Cleaning

3.3.2 Vectorization and Word Embedding

Deep neural networks frequently employ vectorization and word embedding for natural language processing (NLP) applications including sentiment analysis, machine translation, and text categorization.

Text data must be vectorized in order to be transformed into a numerical representation for deep neural network input. The bag-of-words (BoW) model, which represents a document as a vector of word counts, is a popular method of vectorization. Another method is the term frequency-inverse document frequency (TF-IDF) model, which gives each word a weight depending on how frequently it appears in the text and how frequently it appears within the corpus as a whole. Yet, the BoW and TF-IDF models both disregard the meaning and context of the words used in a document. Word embedding has a role in this. Using a method called word embedding, each word is converted into a dense vector in a high-dimensional space, where words that are similar to one another are clustered together.

Word2vec is one of the most popular word embedding methods used in NLP. A neural network technique called Word2vec learns from word spread representations to produce accurate representations without the requirement for labels. When given enough training data, the network generates word vectors with fascinating characteristics. Continuous Bag-of-Words Model (CBOW) and Continuous Skip-gram Model are the two kinds of Word2vec. In essence, CBOW predicts the word based on the provided text, whereas Skip-gram guesses the text based on the provided word.

Using the Word2Vec method, Rehman et al. trained main word embeddings in 2019. The Word2Vc computes the distance between words and groups together similar words after converting the text into a vector of digital values in accordance with the meanings of the terms. The model includes a set of features that were extracted using convolution and global max-pooling layers in addition to long-term dependencies. In order to improve accuracy, the provided model additionally employs dropout function, normalization, and a corrected linear unit. In a different study, constructed Word2Vec models using a corpus obtained from several Arabic-language publications. They used CNN with variation text feature extractions and machine learning techniques on this dataset. The outcome indicated increased accuracy for the sentiment categorization job. An Auto Arabic Lexicon was developed using the best Word2Vec model and used with various ML techniques.

The goal of the Bag of Words (BoW) model is to convert the provided text into a more straightforward numerical representation known as a vector. In order for the computer to interpret the sentence, it is then expressed as a string of numbers. Each vector component also indicates whether a certain word is present or absent. In order to show how frequently a term appears, frequency counts are also used. The model replaces this with a binary representation, either a 1 or a 0. A vector with the same size as the vocabulary is the outcome of BOW. Word embedding, on the other hand, has a vector size that is significantly less than the vocabulary size.

Word embedding thereby addresses (BOW) model's flaws. Similar qualities serve as representations for the parallel texts. Sentiment analysis and text categorization are only two NLP applications that make use of word embedding. The effectiveness and accuracy of sentiment analysis were enhanced in numerous experiments by the use of word embedding. BOW still has limitations even though it is frequently used to evaluate emotion. For instance, inconsistent wording might cause numerous publications to portray the same thing since they use the same phrases. BOW also ignores the semantics of words.

Arabic tweets on violent incidents were gathered and categorized into two categories of sentiments. Their approach is based on Google's Word2vec scheme, which highlights the meaning of words through the use of a deep learning-inspired technique. Weighted average and Word2vec were used to display tweets. Sentiment was also measured using SVM and Random Forest machine learning methods. We have carried out some research to validate our approaches using the cross-validation methodology.

Deep neural networks may be better able to comprehend the meaning of the text if word embeddings are used to capture the semantic links between words. The terms "good" and "great," for instance, would be situated close to one another in the embedding space in a sentiment analysis.

3.3.3 Continuous Bag-of-Words Model (CBOW)

In natural language processing, the Continuous Bag-of-Words (CBOW) model of neural networks is frequently employed for word embedding (NLP). The Word2Vec technique, which uses CBOW as a variation, is made to learn vector representations of words based on their co-occurrence in a sizable corpus of text data.

The CBOW model is a shallow neural network that predicts the target word in the middle of a fixed-length window of context words. An input layer, a hidden layer, and an output layer make up the model. The context words are represented as one-hot vectors in the input layer. The input vectors are first transformed linearly by the hidden layer before being activated by a nonlinear function, such as the hyperbolic tangent (tanh) function. A softmax operation is used in the output layer to create a probability distribution across the vocabulary, where each element reflects the likelihood that the target word will really be that specific word.

To reduce the cross-entropy loss between the predicted probability distribution and the real probability distribution, which is represented as a one-hot vector for the target word, the hidden layer's weights are modified during training using stochastic gradient descent (SGD). Although they are updated, the input layer's weights are not used for prediction.

The weights of the hidden layer can be utilized as word embeddings for further NLP tasks, such as sentiment analysis, machine translation, and text classification, after the CBOW model has been trained. The word embeddings, which are based on the co-occurrence of terms in the training corpus, capture the semantic links between words.

The CBOW model has the benefit of being computationally more effective than alternative neural network architectures for word embedding, such as the Skip-gram model. Word embeddings for frequently occurring words that appear in the training corpus can be learned using the CBOW model.

3.3.4 Skip-Gram (SG) model

In natural language processing, the Skip-Gram (SG) model is a typical neural network design for word embedding (NLP). The Word2Vec approach, which the SG model is a variation of, is made to learn vector representations of words based on their co-occurrence in a sizable corpus of text data.

The SG model is a shallow neural network that predicts a fixed-length window of context words that are present around the target word given as input. An input layer, a hidden layer, and an output layer make up the model. The target word, which is encoded as a one-hot vector, is represented in the input layer. The input vector is transformed linearly by the hidden layer before being activated by a nonlinear function, such as the hyperbolic tangent (tanh) function. One neuron is present in the output layer for each context word in the

window. With the use of a softmax operation, each neuron creates a probability distribution across the vocabulary, where each element denotes the likelihood that the context word will contain that specific word. To increase the likelihood of correctly predicting the context words given the target word, stochastic gradient descent (SGD) is used to adjust the hidden layer's weights during training. Although they are updated, the output layer's weights are not used for prediction. The hidden layer's weights can be utilized as word embeddings for further NLP tasks, such sentiment analysis, machine translation, and text classification, after the SG model has been trained. The word embeddings, which are based on the co-occurrence of terms in the training corpus, capture the semantic links between words.

The SG model has the benefit of being more successful than the Continuous Bag-of-Words (CBOW) model at learning word embeddings for uncommon words that only sometimes appear in the training corpus. The SG model is also good at capturing the syntactic and semantic links between words, such as verb-object and adjective-noun pairings.

3.3.5 Feature Extraction

Tf-idf and word embeddings are the two primary straightforward and issue-independent feature extraction approaches that we used. The term frequency-inverse document frequency (tf-idf term weight) measures how pertinent a word is to each document in a collection of documents. To compare against utilizing word embeddings as features, we solely utilize tf-idf weights with traditional machine learning methods in our tests.

The most common distributed representation of words is word embeddings, which brings us to our second point (or terms). Every word in the vocabulary is represented as a vector with a few hundred dimensions, where words with the same meaning are near together and words with different meanings are far away. Using the situations in which the words are used, one may learn the vector representation of the words. Word2Vec is one of the most widely used methods for effectively learning a solitary word embedding from a text corpus (Mikolov et al., 2013). Skip Gram and Continuous Bag of Words are two alternative instructional approaches for learning the embeddings (CBOW).

The continuous skip-gram takes the current word as input and learns the embeddings by predicting the surrounding words, whereas the CBOW model takes the current word as input and learns the embeddings by predicting the surrounding words (Mikolov et al., 2013). We employed the pre-trained Arabic word embedding model AraVec2.0 (Soliman et al.,

2017) in our experiments. This model offers a variety of pre-trained Arabic word embedding model architectures, each of which is trained on one of three distinct datasets: tweets, web pages, and Wikipedia Arabic articles. Moreover, two models are created for every dataset, one using Skip Gram and the other using CBOW. As we focus on tweets for our research, we employed the pre-trained Skip-Gram 300D-embeddings trained on more than 77M tweets. The pre-trained model was used to both traditional and neural learning techniques. The average vector of all the tweet word embeddings is calculated and utilized as the feature vector of the tweet in order to use it with traditional learning techniques. In contrast, the embedding vectors are employed in neural learning models to set the weights of the embedding layer, which is subsequently linked to the other layers in the network.

3.3.6 Annotation

It is required to annotate the data in order to provide labels for the network to learn from while training a Convolutional Neural Network (CNN) for Arabic text analysis. Annotation is the process of labeling or marking up data to produce a ground truth for machine learning models that can be utilized for training and evaluation. When annotating Arabic literature, certain elements of the text, such as words, sentences, or sentiment, are identified and labeled.

Creating the categories or labels that will be used to categorize the text is the first stage in the annotation process. Sentiment (positive, negative, or neutral), topic (politics, sports, or entertainment), or entity are some examples of these categories (person, organization, location). The annotator can start labeling the text in accordance with the categories after they have been established. Several techniques can be used to annotate Arabic text, depending on the work at hand and the resources at hand. Using human annotators to manually label the data is one such approach. Although this method might be time-consuming and costly, it guarantees excellent annotations and permits the incorporation of subtle descriptors.

Using automated annotation tools like named entity or part-of-speech taggers is another option. To automatically label the text, these solutions make use of pre-existing models and rules, which can save time and effort. Yet the quality of the annotations could be inferior to that of those created by human annotators, and the tools might not be able to handle all kinds of annotation assignments.

The annotated data may then be utilized to train a CNN for Arabic text analysis. In order for the CNN to anticipate what will happen with fresh, unlabeled material, it must learn to identify patterns and characteristics in the text that are connected to the classified categories. It is a popular machine learning approach to carry out this procedure, which is referred to as supervised learning.

Since there is no consensus on what constitutes cyber bullying speech, identifying it can be difficult. The classification of a statement as hate speech is a subjective process based on individual opinion. The following principles were adopted to help human annotators recognize cyberbullying in order to lessen this subjectivity [17,18].

Hence, a post is deemed bullying if it exhibits one or more of the following traits:

1. Using offensive adjectives, terms, or slurs to insult or defame a particular group.
2. Justifying or defending bullying crimes.
3. Supporting and fostering hatred.
4. Promoting one group's superiority over another.
5. Making threats and calling for violence.
6. Disparaging and negative stereotypes.
7. Use of irony and jokes to make fun of and humiliate the target due to a protected attribute.
8. Unique situations: Self-attacking, in which the speaker uses derogatory language to attack a protected attribute of himself.

3.3.7 Training Data and Test Data

A crucial machine learning and data science approach is the division of a dataset into training and test data. With this method, a portion of the data is used to train a model and a different portion is used to assess the model's performance. This approach aims to prevent overfitting, which happens when a model grows overly complicated and performs admirably on training data but badly on test data. Overfitting occurs when a model learns to fit the noise in the training data instead of the underlying patterns, which makes it difficult for the model to generalize to new data.

Choosing how much of the dataset will be utilized for training and how much for testing is the first stage in dividing the dataset. Generally, training uses 70–80% of the data, whereas testing uses the remaining 20–30%. To guarantee that the classes are equally represented in the training and test data, this split can be carried out at random or with the use of a stratified sampling strategy.

Following the data's division, the model is trained using the training set and assessed using the test set. Depending on the nature of the issue and the model's goal, several evaluation metrics are employed to evaluate the model's performance. For instance, standard assessment measures for classification problems include accuracy, precision, recall, and F1 score.

It's vital to remember that the performance of the model on test data rather than training data is a stronger measure of its capacity to generalize to new data. It is crucial to adjust the model's hyperparameters depending on the performance of the test data rather than the performance of the training data.

Creating a deep learning model requires separating a dataset into training and test data. It aids in avoiding overfitting, evaluating the model's effectiveness, and fine-tuning its hyperparameters for improved generalization.

3.3.8 Metrics Evaluation

Accuracy, precision, recall, and F1 score are metrics used to evaluate the performance of a classification model. They are particularly useful in assessing the performance of models that aim to predict binary outcomes, such as spam/not spam, fraud/not fraud, or positive/negative.

Accuracy is the simplest and most commonly used metric for classification models. It measures the percentage of correctly predicted instances among all the instances in the dataset. In other words, accuracy tells us how well the model can classify instances into their respective categories. It is calculated as:

$$\text{Accuracy} = (\text{true positives} + \text{true negatives}) / (\text{true positives} + \text{false positives} + \text{true negatives} + \text{false negatives})$$

Precision is a measure of the proportion of true positives (correctly predicted positive instances) among all the instances that the model predicted as positive. In other

words, it tells us how precise the model is in predicting positive instances. Precision is calculated as:

$$\text{Precision} = \text{true positives} / (\text{true positives} + \text{false positives})$$

Recall is a measure of the proportion of true positives among all the actual positive instances in the dataset. In other words, it tells us how well the model can identify positive instances. Recall is calculated as:

$$\text{Recall} = \text{true positives} / (\text{true positives} + \text{false negatives})$$

F1 score is a measure of the harmonic mean of precision and recall. It is a balance between precision and recall and is useful in cases where we want to consider both metrics equally important. The F1 score is calculated as:

$$\text{F1 score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

A high accuracy score does not necessarily mean that the model is performing well, especially when dealing with imbalanced datasets where the distribution of classes is uneven. For instance, in a dataset where the majority of instances belong to one class, a model that always predicts that class will have high accuracy but poor precision and recall. Therefore, it is essential to consider precision, recall, and F1 score alongside accuracy when evaluating the performance of a classification model.

4 CHAPTER FOUR: DATASET CORPUS

By the use of mobile applications, the internet, and social media portals, there has been a tremendous increase of data in recent years. People are now able to share their opinions about certain topics due to the rapid development of technologies and social media platforms. People all around the world use a lot of these social media sites to express their opinions. By evaluating prior evaluations, it has become simpler to identify cyberbullying thanks to technologies like machine learning (ML) and deep learning (DL). Also, the importance of movie reviews in determining a film's rating and financial success has grown. Also, movie producers might use these critiques to tweak their productions in response to critics' assessments.

Humans cannot, however, analyze all of the data to determine if a person's opinion is favorable or unfavorable due to the vast amount of reviews and information gathered. To improve the production and analysis of the sentiment analysis, we must rely on emerging ML technologies.

4.1 Data Collection

A detailed description of the suggested data gathering methods is discussed in this section. The three Arabic language varieties MSA and Dialect Arabic (DA) are all present in combination in Arabic social media sites. Even if there are some commonalities across the three categories, there are also some significant distinctions that might lead to subpar application of SA. The Python programming language and its libraries are among the most flexible and well-liked techniques used in data analytics, notably for machine learning. The manual technique, which relies on human labor, and the automatic method, which uses an annotation tool, are only two of the various ways to annotate corpora. Figure 12 shows example about the raw data crawled from API Tweeter.

1	Tweet Id	Text	Name	Screen Name	UTC	Created At	Favorites	Retweets	Language	Client	Tweet Type	URLs	Hashtags	Mentions	Media Type	Media URLs
2	1641757141949116	@vill88 احمد عياره عن ديه قاعده على فلي https://t.c.Reno	Reno_832			2023-03-31T:Fri Mar 31 10	0	0	ar	<a href="http	Reply		0	1	photo	https://pbs.twimg.com/media/FoXaoIO0AA
3	16417565774336241	الذك واهيك بالمشحه بالكرشه اقدم كرامتك واهلي رحتك دكتور كرك	uorm554			2023-03-31T:Fri Mar 31 10	0	0	ar	<a href="http	Tweet		13	0	photo	https://pbs.twimg.com/media/FoW5xXoAE
4	1641756503106371	الذك واهيك بالمشحه بالكرشه اقدم كرامتك واهلي رحتك دكتور كرك	uorm554			2023-03-31T:Fri Mar 31 10	0	0	ar	<a href="http	Tweet		13	0	photo	https://pbs.twimg.com/media/FoW5xXoAE
5	1641756009960972	مُتزوج او ديه مُتزوجه واين اشدك تازين خوتك بقوه ؟ برسه Yousef	Dunhill05			2023-03-31T:Fri Mar 31 10	0	0	ar	<a href="http	Tweet		0	0		
6	1641754502993956	@Oge_9 حق اناي ديه تخلفي بيتي بعوزي 9	الجب	2m_87		2023-03-31T:Fri Mar 31 10	0	0	ar	<a href="http	Reply		0	1		
7	1641753771217768	واحد لالطياره جداني اشي عليه وسوئه cpr ويلعوزانيه ويلعوزانيه KindaVenting	KindaVenting			2023-03-31T:Fri Mar 31 10	0	0	ar	<a href="http	Tweet		0	0		
8	1641751379684343	@PaparakaHmadi صدا ديه بدھ 🍌 كبح شدم ، يعني بي	ariaahmadi			2023-03-31T:Fri Mar 31 10	0	0	fa	<a href="http	Reply		0	1		
9	1641749910360932	RT @aa_8600 ب وة فاح حدود بين دولتين	Waleed Alhor	WAlhoraibi		2023-03-31T:Fri Mar 31 10	0	0	ar	<a href="http	Retweet		1	0	video	https://video.twimg.com/ext_tw_video/16
10	1641748534172045	@ahmeddm5 بجة غار صافقه تبول مائل HasanKutbi	1207_bq			2023-03-31T:Fri Mar 31 10	0	0	ar	<a href="http	Reply		0	2		
11	1641746276034260	RT @ali_garwal رفت المحايالي نطن الحب ويذهب	naif707			2023-03-31T:Fri Mar 31 10	0	0	ar	<a href="http	Retweet		4	0	photo	https://pbs.twimg.com/media/Fo8W43wXs

Figure 11: Raw data crawled from Tweeter

4.2 Two-Class Classification

The First dataset which represents a two-class classifications was produced by collecting the Twitter accounts utilizing API Tweeter to find tweets that discussed *cyberbullying and not cyberbullying*. Around 13,000 tweets were collected between Oct 2022 and March 2023; a characteristics of this two-class classification is displayed in Table 2 below. The content of the tweets was extracted in EXEL file then transferred to CSV file after prepossessing the tweets because this format is compatible with a variety of text processing tools. While table 3 presents an example of the collected data for the first dataset that labeled with *Cyberbullying (CB)* and *Not Cyberbullying (NOT-CB)* .

Table 2 : Characteristics Of The Two-Class Classification.

No. of tweets	13481
Start Date	Oct 2022
End Date	March 2023

Table 3: An Example Of The Collected Data For The First Dataset

Tweet	Translate in English	Category
شفنا بس انت اغبي من الغباء نفسه	We saw, but you are dumber than stupidity itself	Cyberbullying
عيش الحياة بسعادة ولا تنكد على نفسك	Live life happily and don't feel bad about yourself	Not Cyberbullying

4.3 Six-Class Classification

The second dataset, which represents the six-class classification, was created by gathering tweets using API Tweeter to look for tweets based on the specified protected features, we further divide the nasty texts discovered at the second level into sex categories as bellow:

- *Sexual cyberbullying (SCB)*
- *Animal phrase cyberbullying (ACB)*
- *Looking cyberbullying (LCB)*
- *Religious cyberbullying (RCB)*

- *Psychological Cyberbullying (PSB)*
- *Not-Cyberbullying (NCB)*

A total of 15,486 tweets were gathered between September 2022 and February 2023; Table 2 below shows the Characteristics of Six-Class Classification.

Table 4: Characteristics of the Six-Class Classification

No. of tweets	15486
Start Date	Sep 2022
End Date	Feb 2023

Table 4 provides instances of real tweets that have been classified as belonging to one of the six cyberbullying categories that previously mentioned.

Table 5: An Example Of The Collected Data For The Six-Class Classification

Tweet	Translate in English	Category
انت حمار بالفطرة بس عندك اجتهاد وتحاول تثبت انك حمار	You are a donkey by nature, but you have a face and try to prove that you are a donkey	Animal Phrase Cyberbullying (ACB)
متخلف صهيوني يهودي	A backward Zionist Jew	Religious Cyberbullying (RCB)
شفتك يا قزم	I saw you, dwarf	Looking Cyberbullying (LCB)
يا فاسقة انتي واهلك	Oh you whore and your family	Sexual Cyberbullying (SCB)
يا تربية الشوارع يا غبي	Oh street breeding you idiot	Psychological Cyberbullying (PSCB)

A total of about 30,0000 tweets were collected using API Tweeter for this study in both the first dataset and the second dataset to detect the cyberbullying in the Arabic social media.

5 CHAPTER FIVE: METHODOLOGY

We approach the identification of cyberbullying as a supervised learning challenge. The essential purpose of the presented methodology is the cyberbullying detection on social platforms which decrease the bullying attitude [65]. This proposed technique can be implemented to assist the cyberbullying fighting and detect the offensive written texts in social media services. Furthermore, the cyberbullying detection on social media may be considered as an effective way to assist and protect the cyberbullying victims [66]. In this work, we tested a number of neural learning models that were trained to find Arabic cyberbullying tweets on Twitter.

The goal of the ongoing study is to provide a framework for reporting cyberbullying that government officials may utilize to make strategic choices. The following are the purposes of the framework's five main modules:

- 1) Automatic tweet extraction
- 2) Data cleaning and pre-processing
- 3) Identification of cyberbullying categories.
- 4) Collecting of the tweets posted by the identified cyberbullying regarding a particular event
- 5) Cyberbullying analysis and reporting (See Fig. 13).

This section illustrates the study methodology that would be implemented to classify the cyberbullying with the proper cyberbullying class. The stages of the presented methodology structure is clarified in Fig. 1, and debated in the following sections.

The methodology of this study is displayed in four stages. In the first stage, the data is collected by utilizing Twitter API and annotated by utilizing the python. In the second stage, the collected data and the annotated data were passed across a data preprocessing and data cleaning to eliminate undesirable symbols and tokens. In the third stage, three deep learning models are implemented to train the data which are a CNN model, a LSTM model and RNN model, while in the final stage is the evaluation which is applied on the models to examine the effectiveness of the proposed techniques.

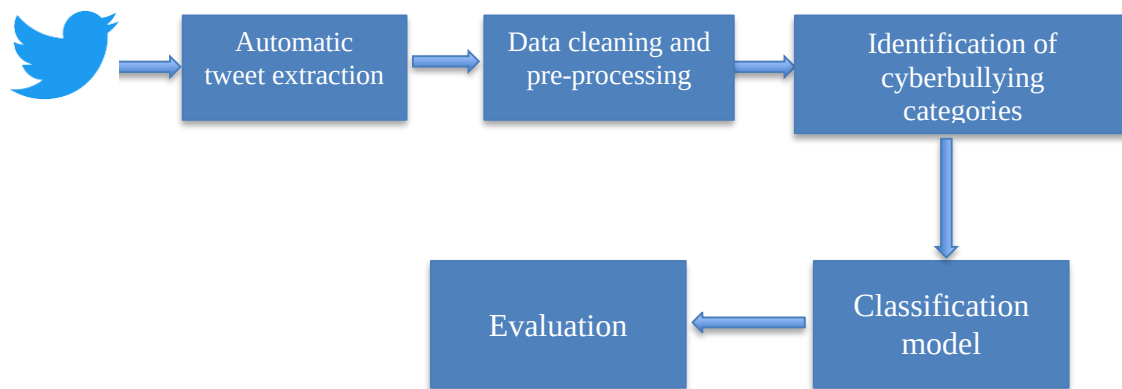


Figure 12: Methodology Structure

5.1 Data Preprocessing Methodology

There are surprisingly few datasets available for the study of online social media forums, presumably as a result of the challenges in developing such datasets under the stringent constraints imposed by social media networks. So, using data collected in two stages, the current research developed two custom databases. In order to identify cyberbullying in each designated issue domain, tweets addressing a variety of specified themes were gathered.

5.1.2 Data Preparation

Data must be preprocessed before being used for emotional analysis. Preparing the dataset and normalizing it is the initial step in preprocessing, which enables the classification algorithm to provide quick and accurate results. These four operations, which make up this stage, can be separated.

- 1- **Transformation:** URLs, hashtags (#), mentions (@), start and final punctuation, lower-case letters in lieu of upper-case characters, and single spaces in place of multiple spaces are all transformed.
- 2- **Tokenization:** Text strings are divided into logical components, such as words or phrases (referred to as "tokens")
- 3- **Filtering:** Search engines are set up to overlook stop words (common terms like the, a, an, in, etc.) both during the indexing of entries for searches and the retrieval of entries in response to search queries, so these words are eliminated.

- 4- **(Lemmatization) Normalization:** This technique, often referred to as "lemmatization," which comes from the word "lemma" or dictionary form, calls for the identification and collection of a single word written in many forms so that it may be studied as a whole.

5.1.1 Collecting Data

For data collection for Arabic social media, Twitter is utilized. The only requirement for the data collection is to register the project in the Twitter developer to gain the consumer key, consumer secret, the access token, and access secret token, which will be utilized in the python code. An Application Program Interface (API) provided by Twitter to assist in collecting the tweets. In this research the data collection was done based on key words which represent the Arabic cyberbullying types. The first dataset was 13481 tweets while the second dataset was 15486 crawled tweets. The collected data extracted as EXCEL file and then transferred them to CSV files.

Utilizing particular keywords is highly beneficial and can detect the bullying tweets based on these keywords enabling labeling the data easily. When getting data from Twitter, the Twitter API contains comprehensive information such as user ID, user screen name, Tweet location, time and the date of tweets, and Tweet text (i.e., the main tweet text containing information about emotions, thoughts, behaviors, and other personally salient information). We used this information to develop a set of features to classify data efficiently from Twitter and use them in different applications. All 30,000 tweets collected for this research related to Arabic swearing words; we use these words as searching seeds in the Twitter search engine.

5.1.3 Training dataset and test data

In this study, two different kinds of datasets were tested. Prior to using sentiment analysis, we had to train the model and produce a machine-readable version via dataset annotation. The process of adding explanations to a corpus collection for NLP purposes is known as annotation. Several annotation techniques are often used depending on the annotators' objectives. For instance, in this study, the first case is with two annotations for calcification (cyberbullying or not cyberbullying). The second annotation was represented with more classification details about each tweets to specify the type of cyberbullying where the second case consists of six annotations for classification which are (sexual cyberbullying, Religious cyberbullying, Animal phrase cyberbullying, looking cyberbullying, psychological cyberbullying and not cyberbullying).

For the Arabic corpus that collected for this study, the data was spitted into training and testing (holdout) with a 70:30 ratio for each case, using the stratified sampling approach, which guarantees an equal class distribution (see Table 6) while table 7 shows the data splitting in both the two cases to training data and test data.

Table 6: Training Data and Test Data

Data	Two-Class Classification		Six-Class Classification					
	CB	Not-CB	SCB	ACB	RCB	LCB	PSCB	NCB
Total data	5929	5278	224	4081	711	969	5380	4126

Table 7: Training Data And Test Data In Two-Class and Six-Class Classifications

DATA	Data in the Two-Class Classification	Data in the Six-Class Classification
Training data	10784	12384
Test data	2696	3096

The flowchart in figure 13 presents the related statistics data of the first and second datasets including the training data and test data in the both cases, while figures 14 &15 show the data splitting to training data and test data in the both two cases. Dataset splitting to training data and test data in with the categories for 2-classes classification case is illustrated in the figures 16 & 17 while the dataset splitting to training and test data with classes for the 6-classes classification case is demonstrated in figure 18.

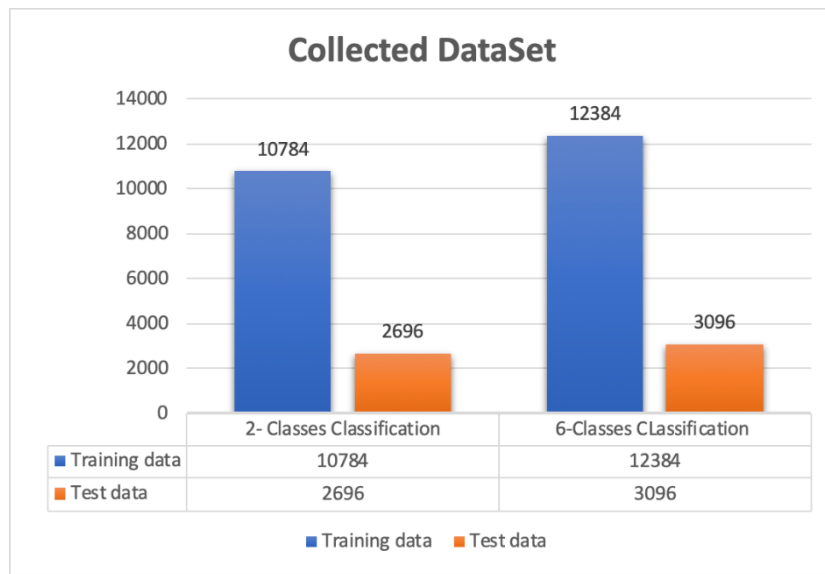


Figure 13: 2- Classes Classification and 6-clses classification Datasets Related Statistics

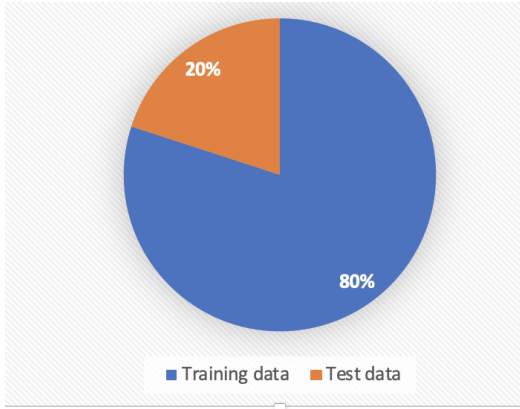


Figure 17: the data set for the 2-Class Classification

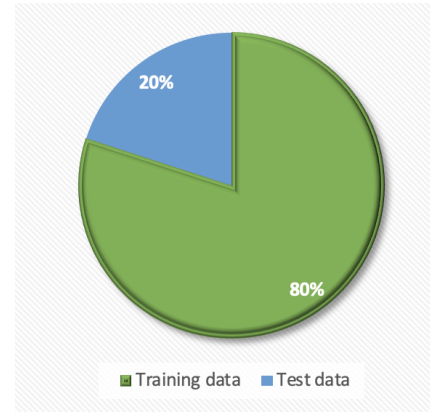


Figure 16: the data set for the 6-Class Classification

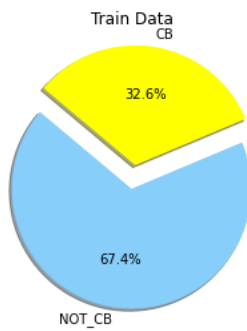


Figure 15: Training data categories for the 6-Class Classification

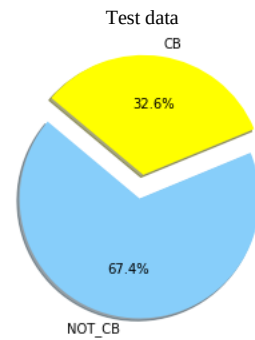


Figure 14: Test Data categories in 6-Classes Classification

Basic text preparation is carried out as follows to get our dataset ready for feature extraction: Punctuation, non-Roman alphabetic and numeric characters (such as user names and URLs), and diacritical marks (such as tashdid, fatha, tanwin fath, damma, tanwin damm, kasra, tanwin kasr, sukun, and tatwil/kashida) are all eliminated. Repeated characters were also eliminated. For more details refer to section for the data preprocessing. The remaining Arabic text has been made normal.

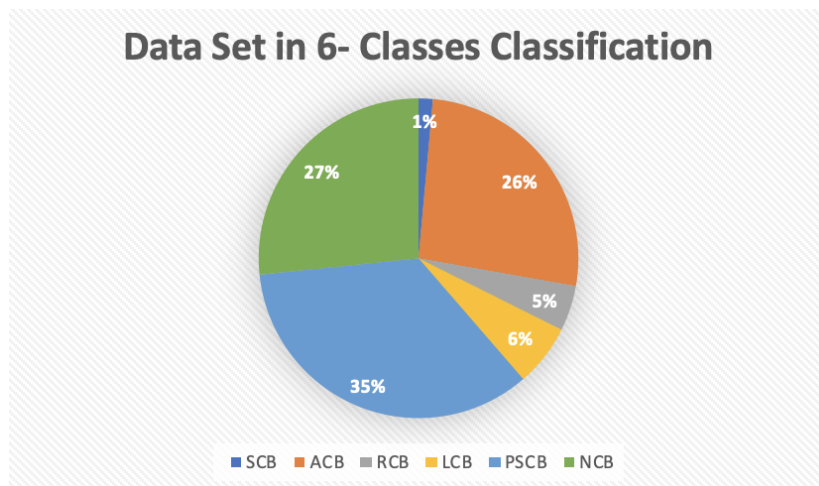


Figure 18: dataset categories for the 6-Class Classification

5.2 Feature Extraction

The pre-trained Arabic word embedding model AraVec2.0 has been used in the experiments. This model offers a variety of pre-trained Arabic word embedding model architectures, each of which is trained on one of three distinct datasets: tweets, web pages, and Wikipedia Arabic articles. Moreover, two models are created for every dataset, one using Skip Gram and the other using CBOW. As we focus on tweets for our research, we employed the pre-trained Skip Gram 300D-embeddings trained on more than 77M tweets. The pre-trained model was used to both traditional and neural learning techniques. The average vector of all the tweet word embeddings is calculated and utilized as the feature vector of the tweet in order to use it with traditional learning techniques. The embedding vectors, on the other hand, are used to set the weights for the embedding layer in neural learning models before it is connected to the other layers in the network.

5.3 Deep learning Models

This study conducted experiments using recurrent neural networks (RNN) and convolutional neural networks (CNN). A variety of RNN designs, including Long Short-Term Memory (LSTM), Bidirectional LSTM (BLSTM), and Gated Recurrent Unit, were tested (GRU). The block diagram of CNN, LSTN and the combined CNN-LSTM are displayed in figures 19, 20, and 21, consequently.

5.4 Evaluation metrics

We provide the total F1-macro attained with various features across all classification levels as we are working with unbalanced data in the 2-class, and 6-class datasets. The confusion matrix and receiver operating characteristic curve (ROC) for the top model are also provided in order to display the area under the curve (AUROC) for the binary classification job and the macro-averaged AUROC for the six class classification tasks. Eq. 1, Eq. 2 and Eq. 3

show the evaluation metrics utilized in this study

$$Precision (P) = \frac{TP}{TP+FP} \quad (1)$$

$$Recall (R) = \frac{TP}{TP+FN} \quad (2)$$

$$F1 - score = \frac{2*P*R}{P+R} \quad (3)$$

Where x represents a class and n represents the total number of classes.

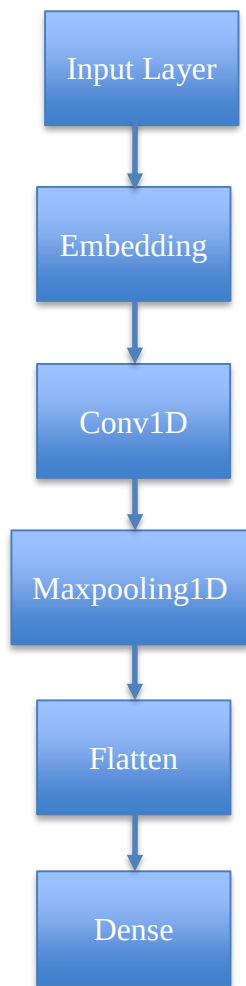


Figure 20: CNN Architecture

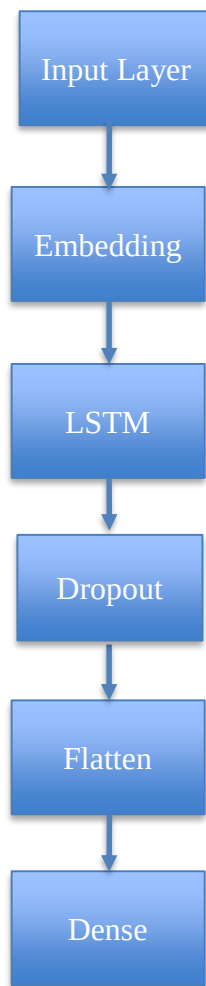


Figure 21: LSTM Architecture

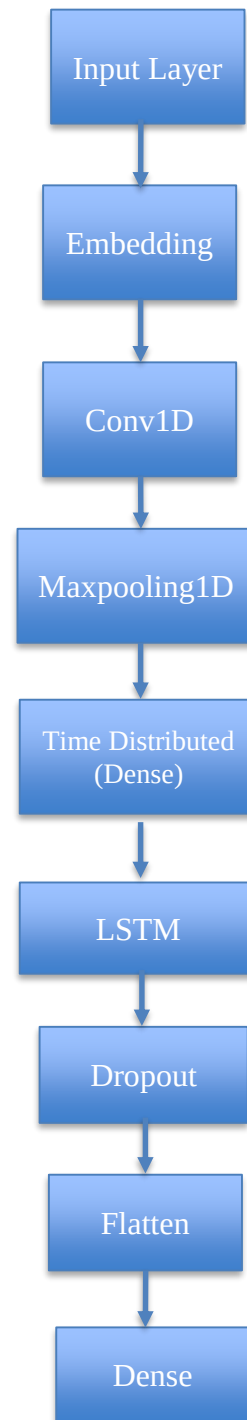


Figure 19: CNN-LSTM Architecture

6 CHAPTER SIX: EVALUATION AND RESULTS

6.1 Experiment Setup

The key elements of our suggested method to identify hate speech in Arabic are presented in this section. The tweets go through a series of preprocessing stages in order to clean up and get the data ready for the training phase (see Section 4). Finally, a number of classification methods are examined for identifying hate speech in Arabic literature. In this study, we assessed three neural network architectures: a CNN-based classifier, a RNN-based classifier, and lastly a classifier that combines CNN and RNN.

6.1.1 Data Preprocessing

We initially pre-processed the training and testing datasets using a number of pre-processing techniques before feeding the pre-processed datasets with tweets as input to the classification models. When working with data that is noisy and informal in nature, like that from Twitter, pre-processing is an essential step. Given the innate ambiguity of the Arabic language and the wide range of dialects utilized in the Twitter, this has even greater significance for the Arabic language. Before supplying our data as an input to our tested models, we performed a number of preprocessing processes on it. See section 5.1.

6.1.2 Deep Learning Neural Network Models

The architectures of our assessed neural network models (CNN, RNN, and CNN + RNN) are described in this section. All models use Word2Vec [38] to express the features as word embeddings; the feature representation is covered in more detail in the next part. After that, the neural network model (CNN, LSTM, or CNN + LSTM) receives these embeddings. The subsections that follow each architecture model's description.

6.1.2.1 Feature Representation

All of the suggested architectures start with an embedding layer that uses a pretrained word2vec model to transfer each tweet, which is represented as a series of integer indices, to a 300-dimension vector space. Using the Continuous Bag of Words (CBOW) training approach, we created our own word2vec model. We used a portion of the data we had gathered as a training collection. 536,000 tokens and 17.6 million tweets in all were included in the dataset. All tweets were preprocessed before to training using the same preprocessing techniques outlined above. Given that tweets are often short, we chose a window size of 3

for the model's hyperparameters. We left the remaining hyper-parameters at their default values and set the vector dimensions to 300. In order to get the tweets ready for the embedding layer, we first built a vocabulary index using word frequency, where each distinct word in our dataset was given a distinct integer value (0 was reserved for padding). In order to make sure that every sequence had the length of the longest tweet in the dataset, we first turned each tweet into a series of integer indexes and padded them. Sequences from the tweets were then sent into an embedding layer, which translated word indexes to previously learned word embeddings. An $(m \times 300)$ tweet matrix, where m is the length of the longest tweet in the dataset, is what the embedding layer's output looked like.

6.1.2.2 CNN Architecture

The first model we assessed was a CNN model for text categorization that was based on Kim's work [39]. Our CNN architecture, as shown in Figure 1, is composed of five layers: the input layer (also known as the embedding layer), the convolution layer, which is essentially made up of three parallel convolution layers with different kernel sizes (2, 3, and 4), the pooling layer, the hidden dense layer, and the output layer. As we previously discussed, the embedding layer maps every tweet into 300-dimensional real valued vectors. The embedding layer then sends an input feature matrix with a shape of $m \times 300$ to a dropout layer with a rate of 0.5, where m is the length of the longest tweet. Dropout layer is primarily used to lessen the issue of overfitting. The three parallel convolution layers then receive the output. A rectified linear unit function for activation and 100 filters with kernel sizes of 2, 3, and 4 are present in each layer. Each layer applies the 100 filters as part of the convolution process, creating 100 new feature maps with a size of $(m - (\text{kernel size} - 1)) \times 100$, which are then fed into a dropout layer with a rate of 0.5. These convolutional features are then fed into a max pooling layer (global), which performs down sampling and generates 3 (1×100) vector outputs. These three vectors are then combined and given as input to a dropout layer with a 0.5 dropout rate, which is followed by a dense layer that is completely linked and contains 50 neurons. The final predictions are then generated by feeding the output to the output layer with sigmoid activation. The summarization of the hyper-parameters of the CNN model is displayed in table 8. The architecture of the CNN is demonstrated in figure 22.

Table 8: Summarize The Hyper-Parameters Of The CNN Model

Hyper-parameters	Value
Number of filters	64
Filter size	3
Dropout layer rate	0.2
Learning rate	0.001
Number of epochs	10
Batch size	50

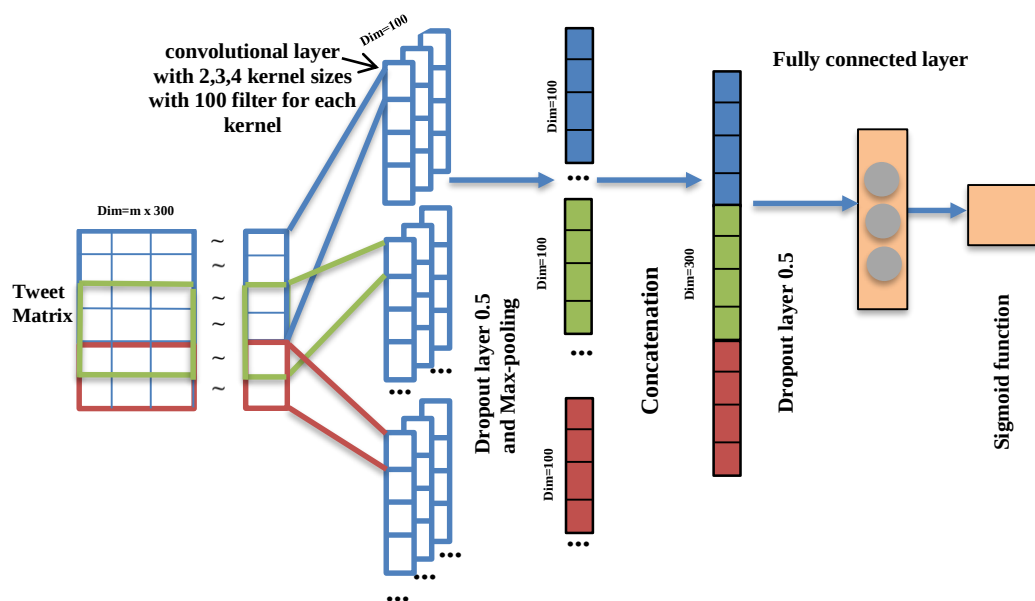


Figure 22: Convolutional neural network (CNN) architecture.

6.1.2.3 Evaluation metrics

The evaluation done by applying the equations 1, 2 and 3 in section 5.4.

6.1.2.4 LSTM Architecture

The table below displays the hyper-parameters for the LSTM model.

For more details about the architecture of the LSTM refer to figure 21 in section 5.3.

Table 9: hyper-parameters for the LSTM model

KJK

Hyper-parameters	Value
Number of hidden unites (RNN)	16
Dropout layer rate	0.5
Learning rate	0.001
Number of epochs	10
Batch size	50

Table 10: Hyper-parameters for CNN-LSTM

Hyper-parameters	Value
Number of filters	64
kernel size	5
Number of LSTM unites	64
Dropout layer rate	0.5
Learning rate	0.001
Number of epochs	10
Batch size	128
pool size	2

6.2 Experiments Results

The outcomes of two different sets of experiments are covered in this section. First up are the in-domain experiments, where every model was developed using the training set of the dataset we built and evaluated using the dataset's testing set.

6.2.1 Results and Discussion

Table 11 lists the outcomes from our tested models for precision, recall, F1-score, accuracy, and cyberbullying recall: CNN, LSTM and CNN + LSTM for the 2- classes classifications. We have some interesting findings. Figure 23 demonstrates the results of CNN model that have been tested. For CB and Not_CB, the figures indicate the same accurate performance patterns measured. Regarding the precision (p), recall and F1, LSTM presented best performance with 96.44%, 97.03% and 96.73%, respectively. Similarly, according to the accuracy, the graph in figure 24 demonstrates that RNN model led to better performance with 94.7% than using CNN or combining the CNN with RNN model.

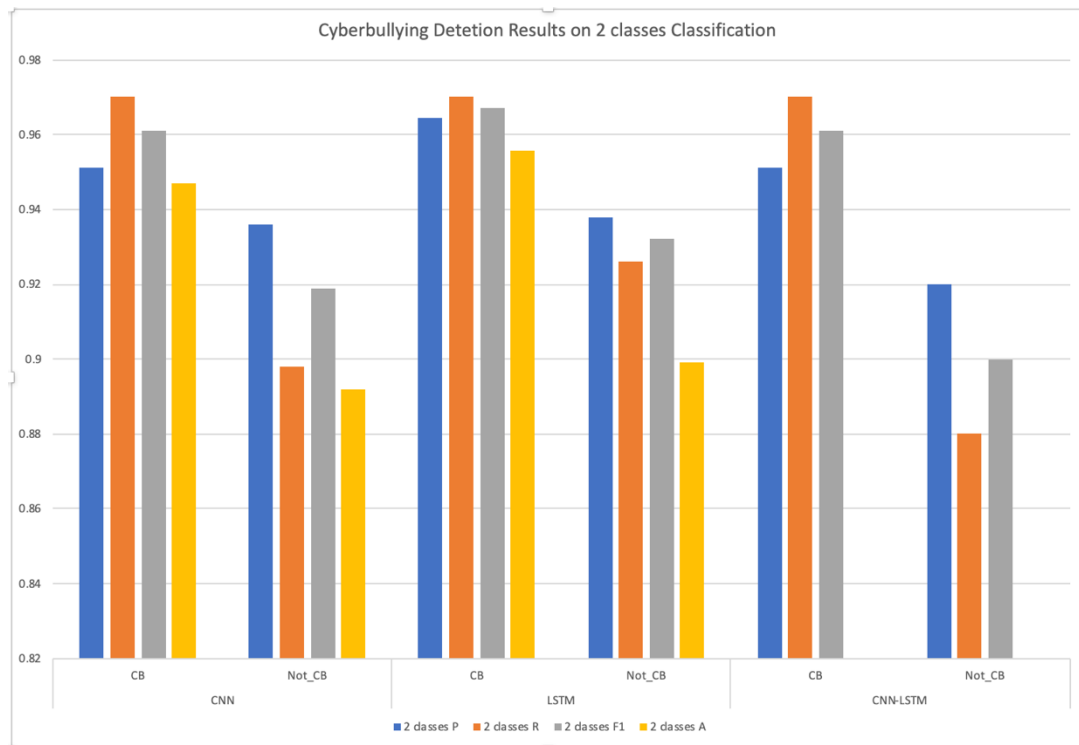
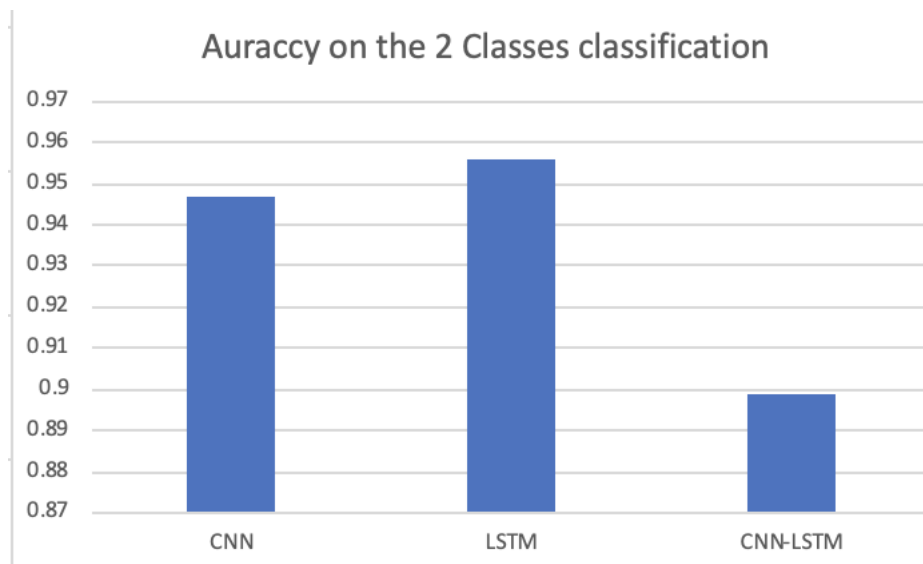


Figure 23: Results of 2 classes classification

Table 11: performance for the 2-classes classification

<i>Model</i>	<i>categories</i>	<i>2 Classes</i>			
		<i>P</i>	<i>R</i>	<i>F1</i>	<i>A</i>
CNN	CB	0.951	0.970	0.961	0.947
	Not_CB	0.936	0.898	0.919	
LSTM	CB	0.9644	0.9703	0.9673	0.9559
	Not_CB	0.9379	0.9261	0.9320	
CNN-LSTM	CB	0.951	0.970	0.961	0.899
	Not_CB	0.92	0.88	0.90	

**Figure 24:** Accuracy on 2 Classes Classification

For the 6 classes of classification, table 12, table 13 and table 14 depict the results from our tested models CNN, LSTM, and CNN + LSTM, sequency. The tables represent the whole 6 classes for precision, recall, F1-score, accuracy, and losses. We have discovered some intriguing things. The tested outcomes of the CNN, LSTM, and CNN + LSTM models are shown in figure 25, figure 26 and figure 27, frequently. The figures show the same precise performance patterns measured for ACB, LCB, NCB, PSCB, RCB and SCB. For CNN model, LCB demonstrated the best performance in terms of precision (p) with 89%, while PSCB showed belter performance for recall (r) with 89%. Repeatedly, LCB gained achieved best results in the recall metrics with 89%. For the F1-score, ACB presented the highest readings with 86%. Similarly, according to LSTM, LCB showed best results with 93% in the (p), while PSCB showed 88%, and LCB presented 89% in F1-sccore. The worst results obtained among the whole classes was from SCB. For the accuracy and the losses in table 13, they are 0.783 and 0.9236, respectively.

Table 12: Results Of CNN On The 6-Classes Classification

<i>Model</i>	<i>Category</i>	<i>6 Classes</i>				
		<i>P</i>	<i>R</i>	<i>F1</i>	<i>A</i>	<i>Loss</i>
CNN	ACB	0.87	0.86	0.86	0.7875	0.984
	LCB	0.89	0.86	0.87		
	NCB	0.78	0.69	0.73		
	PSCB	0.78	0.89	0.83		
	RCB	0.42	0.34	0.38		
	SCB	0.16	0.13	0.14		

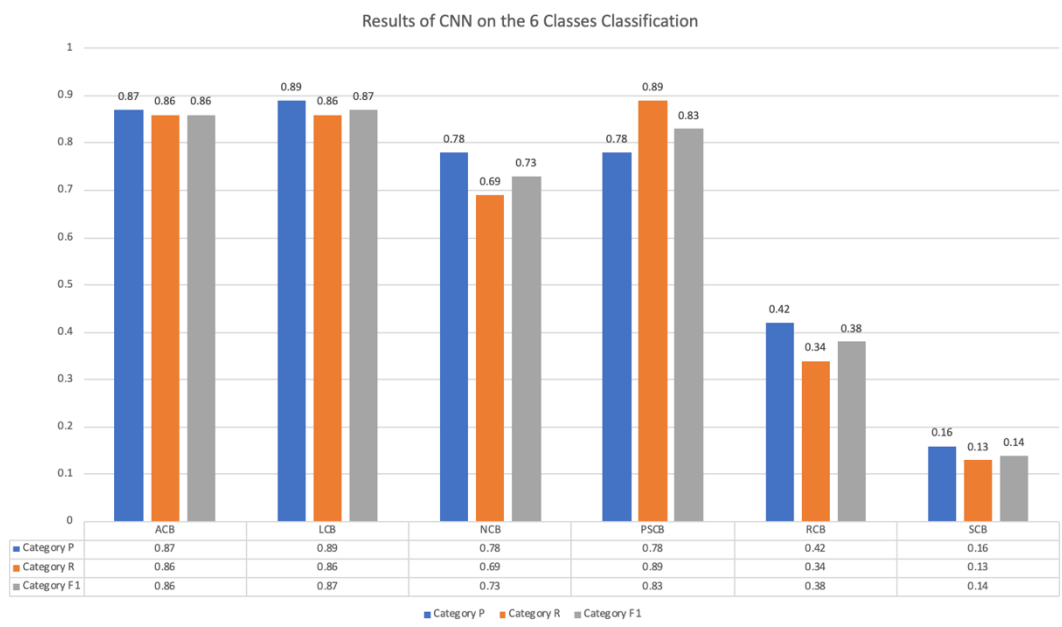


Figure 25: Results of implementing CNN on 6 classes classification

Table 13: Performance Of LSTM On 6 Classes Classification

Model	Category	6 Classes				
		P	R	F1	A	Loss
LSTM	ACB	0.86	0.86	0.86	0.783	0.9236
	LCB	0.93	0.85	0.89		
	NCB	0.75	0.71	0.73		
	PSCB	0.78	0.88	0.83		
	RCB	0.37	0.23	0.28		
	SCB	0.09	0.04	0.06		

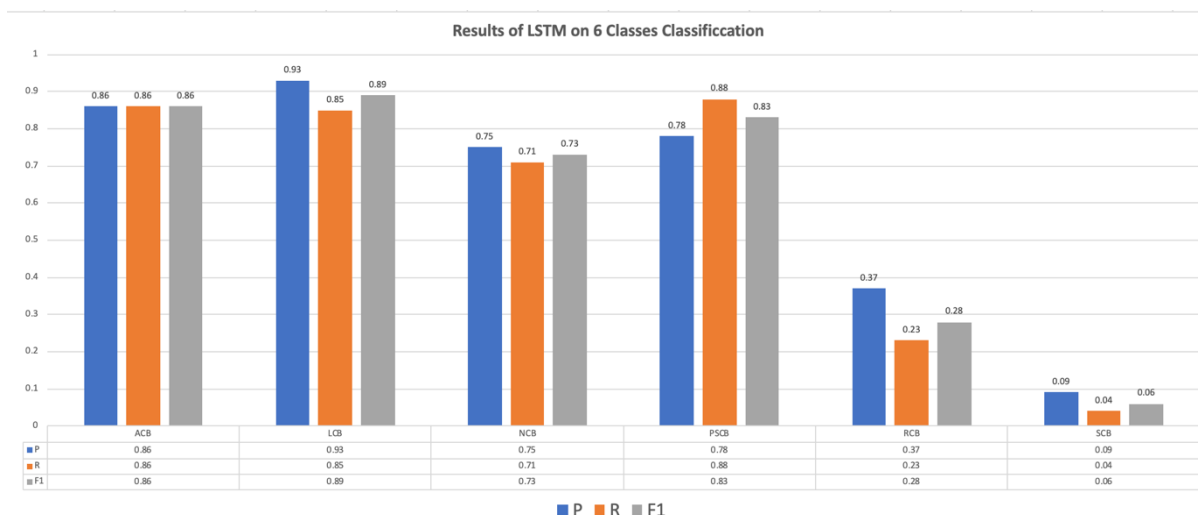


Figure 26: Results of LSTM on 6 Classes Classification

The obtained results in table 14 and figure 29 show performance of the CNN-LSTM model for the 6 classes classification. Category LCB proposed the best results for (p) with 91%, while ACB achieved 90% in recall metrics comparing with the equality in ACB and LCB with 87% for the F1-sccore. The accuracy of the CNN-LSTM is 89.9% while the losses are 1.026.

Figure 28 depicts the Accuracy in CNN, LSTM and CNN-LSTM for the 6 classes classification. It is noticeable that CNN achieved the highest accuracy with 78.75% compared with LSTM 78.3% and CNN-LSTM 77.9%. Regarding the consumed time, CNN

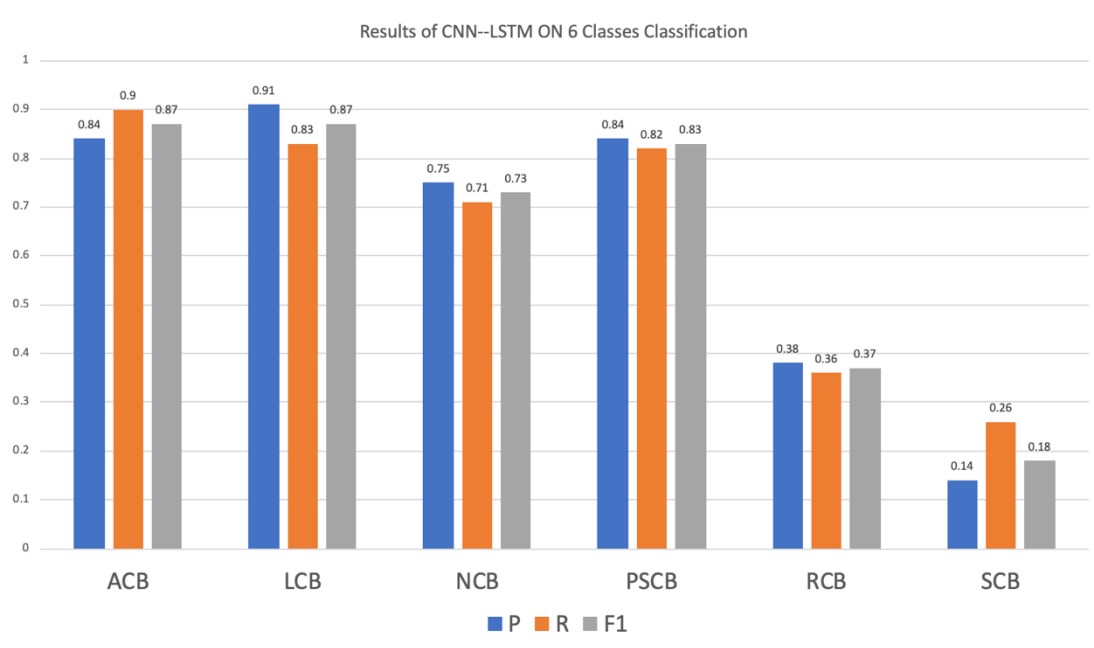


Figure 27: Results of CNN-LSTM on 6 Classes Classification

spent less time in executing the code compared with LSTM. The worst results obtained from SCB and can be improved by improving the datasets.

The results proved that we achieved the target of the thesis where the experiments answered the questions of the thesis.

Table 14: Performance Of The CNN-LSTM on the 6 Classes Classification

Model	Category	6 Classes				
		P	R	F1	A	Loss
CNN-LSTM	ACB	0.84	0.90	0.87	0.779	1.026
	LCB	0.91	0.83	0.87		
	NCB	0.75	0.71	0.73		
	PSCB	0.84	0.82	0.83		
	RCB	0.38	0.36	0.37		
	SCB	0.14	0.26	0.18		

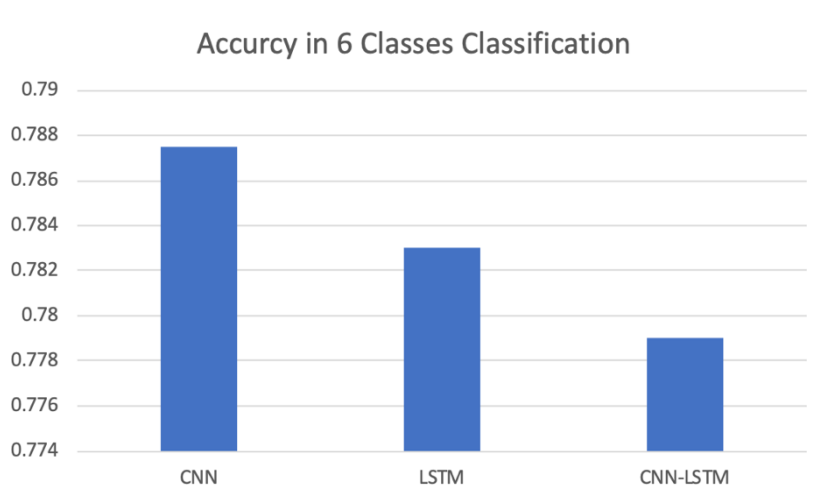


Figure 28: the Accuracy in 6 Classes Classification

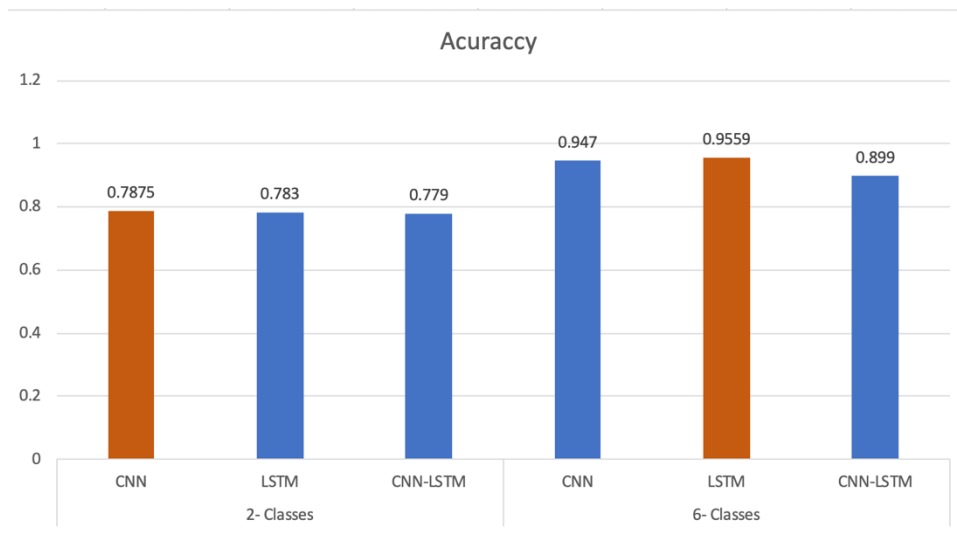


Figure 29: Accuracy in both 2 Classes and 6- Classes classifications

7 CHAPTER SEVEN: CONCLUSION AND FUTURE WORK

The major goals of smart cities are to raise the standard of living of its residents, promote economic expansion while ensuring that development is sustainable, and effectively provide essential public services. Events of a social, economic, or political character prompt a major and immediate public response in a quick-paced, hyper-connected society. As it may be used to anticipate public reactions, public opinion is a form of intelligence that is very important to governments. To effectively and sustainably take action, it is vital to gather and analyze public opinion.

Deep learning models have been successfully used to identify and stop Arabic cyberbullying, with encouraging outcomes. Cyberbullying is an issue that has gotten worse and needs creative solutions as social media usage has grown in the Arab world. Arabic text and speech have been used to recognize and categorize cyberbullying content, and deep learning models like convolutional neural networks and recurrent neural networks have proven to be successful in doing so.

It is crucial to remember that the creation of these models necessitates a sizable amount of labeled data, which might be challenging to come by for Arabic cyberbullying. Accurate model development is further complicated by the cultural and linguistic idiosyncrasies found in Arabic text and voice. Hence, to increase the efficiency and dependability of deep learning models for identifying and preventing Arabic cyberbullying, constant study and collaboration between linguists, psychologists, and machine learning specialists are required.

Due to the growing issue of cyberbullying speech on Arabic Twitter, an efficient automatic cyberbullying detection solution is urgently needed. The methodology suggested in this paper provides real-time sentiment intelligence reports by combining a number of approaches, including automatic data extraction, cyberbullying recognition, sentiment annotation tools, and a hybrid approach to text categorization. In this study, we created a public dataset of about 30,000 tweets to specify the effectiveness of the proposed method. The dataset had been classified into 2 cases, where the first case contains 2-classes classification labeled as Cyberbullying (CB) and Not_Cyberbullying (Not_CB), and the second case consists of 6-classes labeled as: Sexual CyberBullying (SCB), Animal phrase cyberbullying (ACB), Looking cyberbullying (LCB), Religious cyberbullying (RCB),

Psychological Cyberbullying (PSB) and Not-Cyberbullying (NCB). Three distinct models—CNN, LSTM and CNN + LSTM—were assessed and contrasted. The outcomes of our research were encouraging and demonstrated the utility of the suggested models for the detection job. For the 2-classes classification, with 78.7% accuracy, the data demonstrated that CNN outperformed other models. Meanwhile, LSTM proved superior performance with 95.59% in the 6-classes classification.

As a future work, we think there are a number of opportunities to expand and enhance this research. We intend to expand the sorts of data analysis currently being done, such as investigating how attitudes change over time or vary across geographic borders. The dataset may be expanded to include other topics, writing patterns, and writing styles. To improve the detection job outcomes beyond binary classification, the dataset can additionally be annotated with multiple labels. Additionally, we want to concentrate on identifying extreme viewpoints that might serve as a springboard for violent behavior. Lastly, it is planned to enhance the capabilities of our solution by adding support for images and dialect analysis.

REFERENCES

- [1] M. A. Al-Ajlan and M. Ykhlef, "Optimized Twitter Cyberbullying Detection based on Deep Learning," in *2018 21st Saudi Computer Society National Computer Conference (NCC)*, Riyadh, Apr. 2018, pp. 1–5. doi: 10.1109/NCG.2018.8593146.
- [2] D. Mouheb, R. Albarghash, M. F. Mowakeh, Z. Al Aghbari, and I. Kamel, "Detection of Arabic cyberbullying on social networks using machine learning," in *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, 2019, pp. 1–5.
- [3] D. Mouheb, R. Albarghash, M. F. Mowakeh, Z. A. Aghbari, and I. Kamel, "Detection of Arabic Cyberbullying on Social Networks using Machine Learning," in *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, Abu Dhabi, United Arab Emirates, Nov. 2019, pp. 1–5. doi: 10.1109/AICCSA47632.2019.9035276.
- [4] R. R. Dalvi, S. Baliram Chavan, and A. Halbe, "Detecting A Twitter Cyberbullying Using Machine Learning," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, May 2020, pp. 297–301. doi: 10.1109/ICICCS48265.2020.9120893.
- [5] V. Nandakumar, B. C. Kovoov, and M. U. Sreeja, "Cyberbullying Revelation In Twitter Data Using Naïve Bayes Classifier Algorithm.," *International Journal of Advanced Research in Computer Science*, vol. 9, no. 1, 2018.
- [6] A. S. Alqarafi, A. Adeel, M. Gogate, K. Dashitpour, A. Hussain, and T. Durrani, "Toward's Arabic multi-modal sentiment analysis," *Communications, Signal Processing, and Systems*, Jun. 2018, doi: 10.1007/978-981-10-6571-2_290.
- [7] H. AL-Rubaiee, R. Qiu, K. Alomar, and D. Li, "Techniques for Improving the Labelling Process of Sentiment Analysis in the Saudi Stock Market," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, no. 3, Art. no. 3, 30 2018, doi: 10.14569/IJACSA.2018.090307.
- [8] R. Baly *et al.*, "OMAM at SemEval-2017 Task 4: Evaluation of English State-of-the-Art Sentiment Analysis Models for Arabic and a New Topic-based Model," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada, Aug. 2017, pp. 603–610. doi: 10.18653/v1/S17-2099.
- [9] R. Baly *et al.*, "A characterization study of arabic twitter data with a benchmarking for state-of-the-art opinion mining models," in *Proceedings of the third Arabic natural language processing workshop*, 2017, pp. 110–118.
- [10] M. Gridach, H. Haddad, and H. Mulki, "Empirical Evaluation of Word Representations on Arabic Sentiment Analysis," in *Arabic Language Processing: From Theory to Practice*, vol. 782, A. Lachkar, K. Bouzoubaa, A. Mazroui, A. Hamdani, and A. Lekhouaja, Eds. Cham: Springer International Publishing, 2018, pp. 147–158. doi: 10.1007/978-3-319-73500-9_11.

- [11] K. Alomari, H. Elsherif, and K. Shaalan, “Arabic Tweets Sentimental Analysis Using Machine Learning,” 2017, pp. 602–610. doi: 10.1007/978-3-319-60042-0_66.
- [12] S. R. El-Beltagy, T. Khalil, A. Halaby, and M. Hammad, “Combining Lexical Features and a Supervised Learning Approach for Arabic Sentiment Analysis,” in *Computational Linguistics and Intelligent Text Processing*, Cham, 2018, pp. 307–319. doi: 10.1007/978-3-319-75487-1_24.
- [13] A. Elnagar, Y. S. Khalifa, and A. Einea, “Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications,” in *Intelligent Natural Language Processing: Trends and Applications*, K. Shaalan, A. E. Hassanien, and F. Tolba, Eds. Cham: Springer International Publishing, 2018, pp. 35–52. doi: 10.1007/978-3-319-67056-0_3.
- [14] B. Haidar, M. Chamoun, and A. Serhrouchni, “Arabic Cyberbullying Detection: Using Deep Learning,” in *2018 7th International Conference on Computer and Communication Engineering (ICCCE)*, Sep. 2018, pp. 284–289. doi: 10.1109/ICCCE.2018.8539303.
- [15] R. T. Ali and M.-B. Kurdy, “Cyberbullying Detection in Syrian Slang on Social Media by using Data Mining”.
- [16] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching Word Vectors with Subword Information.” arXiv, Jun. 19, 2017. doi: 10.48550/arXiv.1607.04606.
- [17] H. Haddad, H. Mulki, and A. Oueslati, “T-HSAB: A Tunisian Hate Speech and Abusive Dataset,” in *Arabic Language Processing: From Theory to Practice*, Cham, 2019, pp. 251–263. doi: 10.1007/978-3-030-32959-4_18.
- [18] H. Mulki, H. Haddad, C. Bechikh Ali, and H. Alshabani, “L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language,” in *Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy, Aug. 2019, pp. 111–118. doi: 10.18653/v1/W19-3512.
- [19] H. Mubarak, A. Rashed, K. Darwish, Y. Samih, and A. Abdelali, “Arabic Offensive Language on Twitter: Analysis and Experiments,” in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Virtual), Apr. 2021, pp. 126–135. Accessed: Feb. 24, 2023. [Online]. Available: <https://aclanthology.org/2021.wanlp-1.13>
- [20] A. Safaya, M. Abdullatif, and D. Yuret, “KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media,” in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online), Dec. 2020, pp. 2054–2059. doi: 10.18653/v1/2020.semeval-1.271.
- [21] L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, “Hierarchical attention networks for cyberbullying detection on the instagram social network,” in *Proceedings of the 2019 SIAM international conference on data mining*, 2019, pp. 235–243.
- [22] S. Agrawal and A. Awekar, “Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms,” in *Advances in Information Retrieval*, vol. 10772, G.

- Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds. Cham: Springer International Publishing, 2018, pp. 141–153. doi: 10.1007/978-3-319-76941-7_11.
- [23] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep Learning for Hate Speech Detection in Tweets,” in *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, Perth, Australia, 2017, pp. 759–760. doi: 10.1145/3041021.3054223.
- [24] A. Arango, J. Pérez, and B. Poblete, “Hate speech detection is not as easy as you may think: A closer look at model validation (extended version),” *Information Systems*, vol. 105, p. 101584, Mar. 2022, doi: 10.1016/j.is.2020.101584.
- [25] A. S. Saksesi, M. Nasrun, and C. Setianingsih, “Analysis Text of Hate Speech Detection Using Recurrent Neural Network,” in *2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC)*, Bandung, Indonesia, Dec. 2018, pp. 242–248. doi: 10.1109/ICCEREC.2018.8712104.
- [26] S. Biere, “Hate Speech Detection Using Natural Language Processing Techniques”.
- [27] “Cyberbullying Research Center - How to Identify, Prevent, and Respond,” *Cyberbullying Research Center*. <https://cyberbullying.org/> (accessed Jan. 17, 2023).
- [28] S. R and S. Pk, “Cyberbullying: another main type of bullying?,” *Scandinavian journal of psychology*, vol. 49, no. 2, Apr. 2008, doi: 10.1111/j.1467-9450.2007.00611.x.
- [29]: : “STOP Cyberbullying - New Website Launching Soon!! :” <http://www.stopcyberbullying.org/> (accessed Jan. 17, 2023).
- [30] V. Šléglová and A. Cerna, “Cyberbullying in adolescent victims: Perception and coping,” *Cyberpsychology: journal of psychosocial research on cyberspace*, vol. 5, no. 2, 2011.
- [31] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” *Foundations and Trends® in information retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.
- [32] A. Shelar and C.-Y. Huang, “Sentiment analysis of twitter data,” in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2018, pp. 1301–1302.
- [33] T. Mahlangu, C. Tu, and P. Owolawi, “A review of automated detection methods for cyberbullying,” in *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, 2018, pp. 1–5.
- [34] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, “Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network,” *Computers in Human Behavior*, vol. 63, pp. 433–443, 2016.
- [35] M. Munezero, M. Mozgovoy, T. Kakkonen, V. Klyuev, and E. Sutinen, “Antisocial behavior corpus for harmful language detection,” in *2013 Federated Conference on Computer Science and Information Systems*, 2013, pp. 261–265.

- [36] H.-T. Kao, S. Yan, D. Huang, N. Bartley, H. Hosseinmardi, and E. Ferrara, "Understanding cyberbullying on Instagram and Ask. Fm via social role detection," in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 183–188.
- [37] I. Nazar, D.-S. Zois, and M. Yao, "A hierarchical approach for timely cyberbullying detection," in *2019 IEEE Data Science Workshop (DSW)*, 2019, pp. 190–195.
- [38] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1391–1399.
- [39] J. Zhao, K. Liu, and L. Xu, "Sentiment analysis: mining opinions, sentiments, and emotions." MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ..., 2016.
- [40] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093–1113, 2014.
- [41] J.-M. Xu, X. Zhu, and A. Bellmore, "Fast learning for sentiment analysis on bullying," in *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, 2012, pp. 1–6.
- [42] H. Dani, J. Li, and H. Liu, "Sentiment informed cyberbullying detection in social media," in *Joint European conference on machine learning and knowledge discovery in databases*, 2017, pp. 52–67.
- [43] V. Nahar, S. Al-Maskari, X. Li, and C. Pang, "Semi-supervised learning for cyberbullying detection in social networks," in *Australasian Database Conference*, 2014, pp. 160–171.
- [44] A. Seyeditabari, N. Tabari, and W. Zadrozny, "Emotion Detection in Text: a Review." arXiv, Jun. 02, 2018. doi: 10.48550/arXiv.1806.00674.
- [45] S. M. Mohammad, "From once upon a time to happily ever after: Tracking emotions in mail and books," *Decision Support Systems*, vol. 53, no. 4, pp. 730–741, 2012.
- [46] M. J. Vioules, B. Moulahi, J. Azé, and S. Bringay, "Detection of suicide-related posts in Twitter data streams," *IBM Journal of Research and Development*, vol. 62, no. 1, pp. 7–1, 2018.
- [47] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars, "Suicide note classification using natural language processing: A content analysis," *Biomedical informatics insights*, vol. 3, p. BII-S4706, 2010.
- [48] J. A. Patch, "DETECTING BULLYING ON TWITTER USING EMOTION LEXICONS".
- [49] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 8, no. 4, pp. 1–22, 2009.

- [50] A. Abdelali, J. Cowie, and H. S. Soliman, "Arabic Information Retrieval Perspectives".
- [51] N. G. Ali and N. Omar, "Arabic keyphrases extraction using a hybrid of statistical and machine learning methods," in *Proceedings of the 6th International Conference on Information Technology and Multimedia*, 2014, pp. 281–286.
- [52] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, 2005.
- [53] T. Haifley, "Linear logistic regression: an introduction," in *IEEE International Integrated Reliability Workshop Final Report, 2002.*, 2002, pp. 184–187.
- [54] K. Shaalan and H. Raza, "Arabic named entity recognition from diverse text types," in *International conference on natural language processing*, 2008, pp. 440–451.
- [55] A. El-Halees, "Filtering Spam E-Mail from Mixed Arabic and English Messages: A Comparison of Machine Learning Techniques," vol. 6, no. 1, 2009.
- [56] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.
- [57] N. El-Mawass and S. Alaboodi, "Detecting Arabic spammers and content polluters on Twitter," in *2016 Sixth International Conference on Digital Information Processing and Communications (ICDIPC)*, 2016, pp. 53–58.
- [58] R. M. Duwairi, M. Alfaqeh, M. Wardat, and A. Alrabadi, "Sentiment analysis for Arabizi text," in *2016 7th International Conference on Information and Communication Systems (ICICS)*, 2016, pp. 127–132.
- [59] "(PDF) Sentiment Analyzer for Arabic Comments System." https://www.researchgate.net/publication/291748065_Sentiment_Analyzer_for_Arabic_Comments_System (accessed Jan. 21, 2023).
- [60] R. M. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat, "Sentiment Analysis in Arabic tweets," in *2014 5th International Conference on Information and Communication Systems (ICICS)*, Apr. 2014, pp. 1–6. doi: 10.1109/IACS.2014.6841964.
- [61] R. M. Duwairi, "Sentiment analysis for dialectical Arabic," in *2015 6th international conference on information and communication systems (ICICS)*, 2015, pp. 166–170.
- [62] S. Alhumoud, T. Albuhairi, and M. Altuwaijri, "Arabic sentiment analysis using WEKA a hybrid learning approach," in *2015 7th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3K)*, 2015, vol. 1, pp. 402–408.
- [63] A. Al-Zyoud and W. A. Al-Rabayah, "Arabic stemming techniques: comparisons and new vision," in *2015 IEEE 8th GCC Conference & Exhibition*, 2015, pp. 1–6.

- [64] L. S. Larkey, L. Ballesteros, and M. E. Connell, "Light Stemming for Arabic Information Retrieval," in *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, A. Soudi, A. van den Bosch, and G. Neumann, Eds. Dordrecht: Springer Netherlands, 2007, pp. 221–243. doi: 10.1007/978-1-4020-6046-5_12.
- [65] "Bullying and Emotional Abuse in the Workplace: International Perspectives in Research and Practice - ProQuest." <https://www.proquest.com/openview/a9dad329b4fee4a986a040dbba2b5335/1?pq-origsite=gscholar&cbl=36693> (accessed Feb. 13, 2023).
- [66] M. Dadvar and F. De Jong, "Cyberbullying detection: a step toward a safer internet yard," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 121–126.
- [67] J. Escartín, D. Zapf, C. Arrieta, and Á. Rodríguez-Carballeira, "Workers' perception of workplace bullying: A cross-cultural study," *European Journal of Work and Organizational Psychology*, vol. 20, no. 2, pp. 178–205, Apr. 2011, doi: 10.1080/13594320903395652.