

**ANALYZING LARGE LANGUAGE MODELS FOR
CLASSIFYING SEXUAL HARASSMENT STORIES WITH
OUT-OF-VOCABULARY WORD SUBSTITUTION**

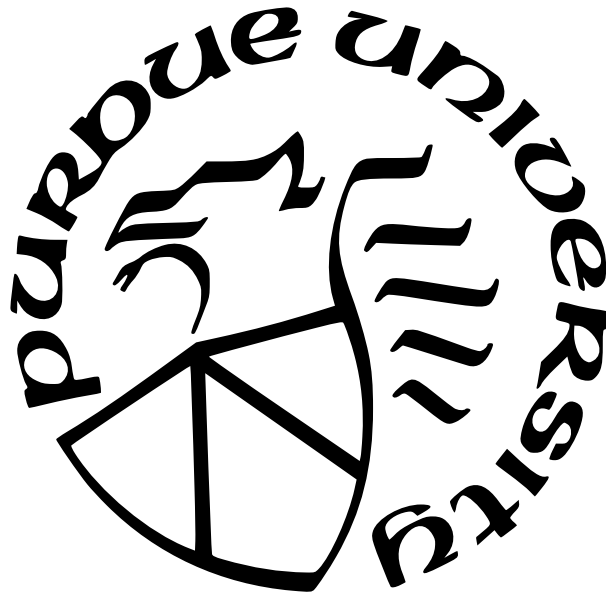
by
Seungyeon Paik

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science



Department of Computer and Information Technology

West Lafayette, Indiana

May 2024

**THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL**

Dr. Tatiana Ringenberg, Chair

Department of Computer and Information Technology

Dr. Julia Rayz

Department of Computer and Information Technology

Dr. Tianyi Li

Department of Computer and Information Technology

Approved by:

Dr. Chad Laux

ACKNOWLEDGMENTS

I would like to express my gratitude to my chair professor Dr. Tatiana Ringenberg for her patient guidance and support. I also wish to gratefully acknowledge my thesis committee, Dr. Julia Rayz and Dr. Tianyi Li for their insightful comments and guidance and my family for their support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES	7
LIST OF FIGURES	8
ABBREVIATIONS	9
ABSTRACT	10
1 INTRODUCTION	11
1.1 Definitions	12
1.2 Scope	12
1.3 Significance	13
1.4 Research Question	13
1.5 Assumptions	14
1.6 Limitations	14
1.7 Delimitations	16
1.8 Summary	16
2 REVIEW OF RELEVANT LITERATURE	17
2.1 Sexual Harassment	17
2.1.1 Definition	17
2.1.2 Impacts	19
2.1.3 Sexual Harassment Dataset	20
2.1.4 Classification of Sexual Harassment Type	21
2.2 Large Language Models(LLMs)	23
2.2.1 Definition and Evolution of LLM	23
2.2.2 Applications of LLM	26
2.3 Out-of-Vocabulary(OOV) Word	28
2.3.1 Definition	28
2.3.2 OOV Word Handling	28

2.4	Summary	30
3	METHODOLOGY	31
3.1	Study Design	31
3.2	Dataset	31
3.2.1	Dataset Description	31
3.2.2	Dataset Preprocessing	33
3.3	Models	33
3.4	OOV Replacement	35
3.5	Classification	37
3.5.1	Dataset splitting and Cross-validation	38
3.5.2	Model Training	39
3.6	Result Analysis	39
3.6.1	OOV Feature Analysis	39
3.6.2	SHAP Analysis	40
3.7	Summary	41
4	RESULTS	42
4.1	Evaluation Metrics	42
4.2	SHAP Analysis	43
4.3	Comparison of Model Prediction on the Original Dataset and Replaced Dataset	45
4.3.1	Same Results Before and After OOV Replacement (C-C, I-I)	46
4.3.2	Different results before and after replacement (C-I, I-C)	49
4.4	OOV Feature Analysis	50
4.5	Summary	51
5	DISCUSSION, LIMITATIONS AND FUTURE PLAN	52
5.1	Discussion	52
5.1.1	Model Performance	52
5.1.2	SHAP Analysis	53

5.1.3	Comparison of Model Prediction on the Original Dataset and Replaced Dataset	54
	Same Results Before and After OOV Replacement (C-C, I-I)	54
	Different results before and after replacement (C-I, I-C)	57
5.1.4	OOV Feature Analysis	59
5.2	Limitations	59
5.3	Future Plan	60
5.4	Summary	60
REFERENCES		61

LIST OF TABLES

3.1	Example of SafeCity dataset	32
3.2	Dataset Distribution	33
4.1	Model Performance on Original Dataset	43
4.2	Model Performance on Replaced Dataset	43
4.3	The most influential words for model prediction (Original dataset)	43
4.4	The most influential words for model prediction (Replaced dataset)	44
4.5	Comparison of Predicted Classes Before and After OOV Replacement (I-I) . . .	47
4.6	Top Common Keywords Before and After OOV Replacement When Correctly Classified	49
4.7	Top Common Keywords Before and After OOV Replacement When Incorrectly Classified	49
4.8	Number of Instances With the Same Influential Word (C-I, I-C)	50
4.9	Domain of OOV words in the Original Dataset	50
4.10	The most frequent OOV words	50

LIST OF FIGURES

3.1	BertForSequenceClassification Model Structure	35
3.2	Proposed OOV word replacement flow	38
3.3	SHAP Analysis Saliency Plot	40
3.4	SHAP Analysis Bar Graph	41
4.1	Analysis of Correct and Incorrect Classifications Before and After Replacement .	45
4.2	Comparison of Consistency in Outcome Between Most Influential Words Selected Before and After OOV Replacement	47
4.3	Comparison of Predicted Classes Before and After OOV Replacement (I-I) . . .	48
5.1	Instance classified correctly both before and after replacement	55
5.2	Instance classified correctly both before and after replacement	55
5.3	Instance classified correctly both before and after replacement	55
5.4	Instance classified incorrectly both before and after replacement	56
5.5	Instance classified incorrectly both before and after replacement	56
5.6	Instance classified correctly before but incorrectly after replacement	57
5.7	Instance classified correctly before but incorrectly after replacement	57
5.8	Instance classified incorrectly before but correctly after replacement	58

ABBREVIATIONS

BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional LSTM
CNN	Convolutional Neural Network
CNN-BiLSTM	Convolutional Neural Network-Bidirectional LSTM
GPT	Generative Pre-trained Transformer
LLaMA	Large Language Model Meta AI
LLM	Large Language Model
LSTM	Long Short-Term Memory
MLP	Multilayer Perceptron
NLP	Natural Language Processing
SVM	Support Vector Machine
T5	Text-to-Text Transfer Transformer
TF-IDF	Term Frequency-Inverse Document Frequency
ULMFit	Universal Language Model Fine-tuning
XLNet	eXtreme MultiLingual Language Model

ABSTRACT

Sexual harassment is regarded as a serious issue in society, with a particularly negative impact on young children and adolescents. Online sexual harassment has recently gained prominence as a significant number of communications have taken place online. Online sexual harassment can happen anywhere in the world because of the global nature of the internet, which transcends geographical barriers and allows people to communicate electronically. Online sexual harassment can occur in a wide variety of environments such as through work mail or chat apps in the workplace, on social media, in online communities, and in games [1]. However, especially for non-native English speakers, due to cultural differences and language barriers, may vary in their understanding or interpretation of text-based sexual harassment [2]. To bridge this gap, previous studies have proposed large language models to detect and classify online sexual harassment, prompting a need to explore how language models comprehend the nuanced aspects of sexual harassment data. Prior to exploring the role of language models, it is critical to recognize the current gaps in knowledge that these models could potentially address in order to comprehend and interpret the complex nature of sexual harassment.

The Large Language Model (LLM) has attracted significant attention recently due to its exceptional performance on a broad spectrum of tasks. However, these models are characterized by being very sensitive to input data [3], [4]. Thus, the purpose of this study is to examine how various LLMs interpret data that falls under the domain of sexual harassment and how they comprehend it after replacing Out-of-Vocabulary words.

This research examines the impact of Out-of-Vocabulary words on the performance of LLMs in classifying sexual harassment behaviors in text. The study compares the story classification abilities of cutting-edge LLM, before and after the replacement of Out-of-Vocabulary words. Through this investigation, the study provides insights into the flexibility and contextual awareness of LLMs when managing delicate narratives in the context of sexual harassment stories as well as raises awareness of sensitive social issues.

1. INTRODUCTION

Online sexual harassment has become a growing problem in modern society as a result of people spending a greater amount of time online. According to the Pew Research Center, 41% of Americans have personally experienced some form of online harassment [5]. Online sexual harassment consists of distinctive characteristics that make it easier for offenders to victimize individuals regardless of their location, target multiple victims simultaneously, avoid being caught, and maintain a certain level of anonymity [6], [7]. In addition, online sexual harassment can happen at any time and anywhere because people constantly use the Internet, and the Internet has the characteristic of allowing people to communicate beyond geographic and time limits. There are many different settings where online sexual harassment can happen, including private social networks, online forums, work email or messengers, and in-game [1], [8]. Therefore, anyone can become a victim of online sexual harassment, regardless of the target’s age or location. Previous studies show that adolescents are more vulnerable to online sexual harassment than adults [9]. Moreover, people who are not familiar with the culture or language may often find it difficult to understand or recognize instances of sexual harassment, even if they do occur [10].

The emergence of the Large language model (LLM) has opened up opportunities for natural language processing (NLP), showing themselves to be useful in a wide variety of areas such as machine translation, text generation, and speech processing [11], [12]. Especially, LLMs are increasingly utilized for solving social issues, including the detection and the online classification of sexual harassment, and NLP demonstrated its broad social impact [13]. Numerous previous studies have been conducted to address the sexual harassment issue, such as sexual harassment detection in conversations or classification of sexual harassment types on social media [14]–[17]. However, the ability to identify and produce consistent results for similar texts is crucial to demonstrate the reliability of LLMs for complex social tasks, especially in sensitive domains like sexual harassment. Due to the nature of LLM, which is highly sensitive to input, prompt sensitivity of LLMs in various linguistic contexts in certain domains is of great interest [3], [18], [19]. It begs the question of whether language models are capable of picking up on the subtleties that are present in language when it comes

to specific and narrow domains such as sexual harassment, given how difficult it is to discern these nuances.

This study aims to evaluate the classification capabilities of LLMs for both original sentences and sentences where OOV words have been replaced. The author also examines the effectiveness of LLMs in categorizing narratives from the sexual harassment story dataset. The author’s goal is to compare the ability of LLMs to comprehend original sentences with OOV words and sentences without OOV words and to explore their capacity to classify narratives accurately within this sensitive domain. Through this study, the author seeks to provide insights into the performance and adaptability of LLMs in this context.

In this research, the author explores the potential of multiple LLMs by employing a SafeCity dataset provided by [20] which is composed of sexual harassment stories. This study’s main goal is to evaluate and compare how well different Large Language Models (LLMs) do at classifying the text containing OOV words and the text without OOV words within the sensitive domain of sexual harassment narratives.

1.1 Definitions

In the broader context of thesis writing, the author defines the following terms:

(commonly: *LLMs*) An artificial intelligence tool that is capable of handling a variety of natural language processing tasks such as question-and-answer, text generation, and machine translation tasks. [21]

(commonly: *NLP*) Computational approaches to analyzing and understanding the text, to obtain human-like language processing in various tasks. [22]

(commonly: *OOV*) Words that are not seen in the training data. [23]

1.2 Scope

This study aims to assess the impact of small differences in input on an LLM’s ability to classify sexual harassment stories before and after replacing Out-of-Vocabulary words.

The scope of this research is the evaluation of LLM’s perception before and after replacing OOV words in the dataset, particularly within the sensitive domain of sexual harassment narratives. In this study, the author will examine the BERT-based classification model to gain comprehensive insights into its proficiency in comprehending datasets with and without OOV words. Our investigation will use a slightly modified multi-level OOV replacement technique proposed by [24] as well as a qualitative analysis of classification results before and after replacing OOV words in the dataset.

1.3 Significance

This research delves into the nuanced and challenging domain of the classification of text within the context of narratives related to sexual harassment. By focusing on the classification of sexual harassment stories and examining the performance of LLMs, this study serves as a valuable resource for understanding the capabilities of LLMs in a sensitive context. This approach will also enable us to explore how LLM handles slightly different inputs in sensitive domains that may have niche vocabulary and concepts. By situating this research in the context of sexual harassment, it advances the larger goal of addressing and promoting awareness of this important issue. If LLM showed robustness to small differences, it might be employed to identify instances in which non-native English speakers are uncertain about online sexual harassment situations. Furthermore, the LLM features towards the sexual harassment domain discovered in this study could contribute to protecting non-native English speakers from online sexual harassment. This study highlights the interdisciplinary value of Artificial Intelligence (AI) in addressing societal challenges and encourages the responsible use of NLP in situations with real-world, significant social repercussions.

1.4 Research Question

The main research question of this study is: How does the replacement of OOV words impact the classification performance of LLMs in the sexual harassment domain? The objective of this question is to compare how LLM perceives scenarios containing and not containing

various OOV words in a narrow domain. By answering this query, the author hopes to shed light on the reliability and sensitivity of LLM’s understanding of OOV words.

1.5 Assumptions

The assumptions for this study include:

- First, the author assumes the selected dataset contains a representative and comprehensive sample of the language used in real-world situations. As shown in previous studies [25]–[27], data quality is a pivotal factor in the success of natural language processing (NLP) models. The author assumes the selected dataset encapsulates the abundance and diversity of language use in the context of sexual harassment, ensuring that the model is exposed to a wide range of language nuances, expressions, and parsing patterns. Representative datasets are considered essential to cultivating the ability of large language models (LLMs) to effectively identify and generalize paraphrasing patterns by capturing the complexity of language specific to sexual harassment narratives. This assumption becomes particularly important given the nuanced nature of the subject in which nuanced changes in language play an important role in communicating meaning and intention.
- Second, the author assumes that the method used for replacing out-of-vocabulary (OOV) words with in-vocabulary (IV) words effectively captures the semantic context of the original sentence. This assumption is based on the premise that the replacement words are semantically similar to the OOV words they replace, thereby preserving the overall meaning and intent of the sentence. Additionally, the author assumes that the replacement process maintains syntactic correctness and grammatical coherence to ensure the fluency and readability of the modified sentences.

1.6 Limitations

The limitations for this study include:

- First, biases in dataset and models. This research acknowledges the possibility of bias in both the dataset and the pre-trained models. Due to the nature of the sexual harassment domain, there are not many datasets available. Also, the author only used the SafeCity dataset for this research. The acknowledgment includes the potential for over- or underrepresentation of particular demographics, viewpoints, or experiences related to sexual harassment in the dataset. The study acknowledges that these biases within the dataset may restrict the applicability of findings to a broader domain because the training data for the models might not adequately represent the wide range of experiences and expressions. In addition, this study recognizes the possibility of biases present in the pre-trained models. Large volumes of random data are exposed to models during the pre-training phase, which may unintentionally contain societal biases found in online content. This study acknowledges that these model biases may affect how the models interpret and handle paraphrased sentences.
- Second, the author focuses exclusively on single-class data instances, excluding multi-class data instances from the analysis. The decision to re-label the dataset to contain only single-class data instances may introduce biases and limitations in the model’s training and evaluation process. By excluding multi-class instances, the study overlooks potentially valuable information and patterns present in the data, limiting the generalizability of the findings. Furthermore, the exclusion of multi-class instances may impact the model’s ability to handle real-world scenarios where text instances may belong to multiple categories simultaneously. Therefore, the findings of this study should be interpreted with caution, recognizing the inherent limitations imposed by the exclusion of multi-class data instances from the analysis.
- Third, the size of the dataset. After selecting single-class data instances, there are 5,002 instances in the dataset. Small dataset sizes may provide limited information for LLMs to understand and categorize sexual harassment stories. Because of this, LLMs could not fully comprehend the various facets and contexts of sexual harassment stories. These limitations due to the small dataset size could limit the generalizability of our findings and reduce the performance and reliability of our models.

1.7 Delimitations

The delimitations for this study include:

- First, this research exclusively investigates the effect of out-of-vocabulary (OOV) replacement on the classification performance of LLM and does not explore alternative methodologies for enhancing model performance, such as data augmentation or feature engineering. While OOV replacement represents one approach to address vocabulary gaps and improve model robustness, it is not the only strategy available for optimizing LLM performance. By focusing solely on OOV replacement, this study may overlook the potential benefits offered by other techniques, such as generating synthetic data through data augmentation or extracting informative features through engineering. Consequently, the findings of this research may provide an incomplete understanding of the broader landscape of methodologies for enhancing LLM performance, emphasizing the importance of future studies that explore and compare multiple strategies in tandem.
- Second, the exclusion of specific LLMs. Recognizing the vast landscape of LLMs, this study deliberately narrows its focus to the BERT-based model. While BERT is widely used in text classification tasks, the exclusion of other models is acknowledged as a deliberate limitation. Focusing on a specific model allows for more detailed comparisons and investigations, even if not all of the models that exist can be explored.

1.8 Summary

This chapter provided the scope, significance, research question, assumptions, limitations, delimitations, definitions, and other background information for the research project. The next chapter provides a review of the literature relevant to sexual harassment, OOV words, and large language models.

2. REVIEW OF RELEVANT LITERATURE

This chapter provides a review of the literature relevant to Sexual harassment, the Large Language Model, and Out-of-Vocabulary words.

2.1 Sexual Harassment

2.1.1 Definition

Sexual harassment is characterized by inappropriate behavior with a sexual component, as defined by [28]. Numerous approaches exist for defining sexual harassment both from legal and psychological perspectives. The legal definition of sexual harassment in the U.S. is provided by The Equal Employment Opportunity Commission (EEOC) Guidelines. They define sexual harassment as, verbal or physical conduct of a sexual nature that unreasonably interferes with the employees work or creates an intimidating, hostile or offensive working environment [29, p. 1]. This definition targets the work environment. In contrast, there are psychological definitions that emphasize the behavior of offenders and the experience of victims [30]. According to [31], sexual harassment is defined as actions that degrade, condemn, or denigrate a person because of their gender. [32] define sexual harassment as unwanted male conduct that prioritizes a woman’s sexual role over her duty as an employee. [33] introduced a comprehensive framework for defining sexual harassment, encompassing six distinct behaviors. These encompass verbal harassment or abuse, subtle coercion towards sexual activities, unwarranted physical contact such as patting or pinching, persistent bodily contact with another person, requests for sexual favors coupled with implicit or explicit threats regarding an individual’s job security, and requests for sexual favors coupled with implicit or explicit promises of favorable treatment in relation to an individual’s employment status. [34] define sexual harassment as unwanted sexual behavior at work that the recipient finds insulting, overwhelmed, or endangers her health. A relatively recent definition from [35] is behavior that is unwanted and that is intended to be offensive, confrontational, intimidating, or humiliating.

The emergence of the digital age has changed the nature of sexual harassment in recent times, raising the pressing problem of online sexual harassment. According to [36], online harassment affects 40% of internet users, with varied degrees of intensity. Compared to other demographic groups, young adults are most likely to encounter online harassment. Approximately 64% of adults under 30 have encountered some kind of online harassment [5].

Conventional interpretations of sexual harassment have mainly focused on physical interactions that occur in public or workplace settings. However, defining, comprehending, and dealing with sexual harassment now presents a number of additional difficulties and complexities due to the digital age. A wide range of actions falls under the umbrella of cyber sexual harassment [37], from overt and unwanted sexual advances to more covert forms of coercion through the internet, like stalking [38], sharing private information without consent, and online grooming [39]. Because of the anonymity and worldwide reach of the internet, offenders can target victims anywhere in the world, making it a widespread problem with far-reaching effects [7], [40].

Researchers define online harassment in various ways. The broad definition is hostile behavior occurs online [41], [42]. [43] have defined online sexual harassment as behavior including sexual requests, image-based harassment, sexual coercion, and hate speech. [44] defined online harassment as inappropriate conduct sent to a young person online or published online for public viewing. [45] define online harassment as aggressive behavior or actions of interpersonal hostility that are conveyed over the internet or other electronic media. They conducted a thorough review of the literature and provided definitions for online sexual harassment, conceptualizing them by bullying components such as intended harm [46]–[48], power imbalance, and frequency of harassment behavior [49]. [46] defines online harassment as an interpersonal activity through a computer to intentionally hurt another worker in the workplace. [49] defines online harassment as frequent online communication that targets a specific person and results significant mental stress and/or the risk of physical harm. [50], [51] conceptualize online harassment as workplace harassment, defining it as actions that the victim believes make the workplace unpleasant or hostile. [52] provides a scoping review of online sexual harassment literature, and attempts to define terms used to describe online sexual harassment in the adolescent population. In this study, the author defines online sex-

ual harassment as the act of causing unwanted sexual discomfort and aversion online such as verbally or visually aggressive sexual conversations, jokes, and insults.

2.1.2 Impacts

According to clinical observations by [53], [54], sexual harassment can affect victims in various aspects, both mentally and physically. From therapeutic experience data and sexual employment survey, [54] found that victims go through certain stages of emotional changes including confusion, self-blame, fear, anxiety, depression, anger, and disillusionment. They also compared sexual harassment victims to victims of rape, battering, and incest, and found all of them experience grief, shame, guilt, fear, and rage. Among the victims of various violence, sexual harassment victims usually experience these emotions for a longer time. [55] analyzed the data from questionnaires and records from the Institute’s crisis counseling service, and found that sexual harassment victims experience psychological stress symptoms such as general tension or nervousness, persistent anger, and fear, as well as decreased productivity and self-confidence. [56], [57] studied the harmful effects of sexual harassment and came to the conclusion that sexual harassment can and does cause serious issues with women’s mental health. In addition, sexual harassment was investigated to have a physical effect on the victim. [58] investigated previous studies [54], [55] and found that sexual harassment victims suffered physical symptoms such as fatigue, migraines, weight loss, and insomnia. Research has shown that teenagers, who use the internet more frequently than adults, are particularly susceptible to online sexual harassment [36], [59]. Studies conducted by [37], [60]–[63] have also shown that adolescents who experience online sexual harassment may experience emotional and physical health issues as a result, with women being the victims more frequently. [64] conducted an online survey on 594 adolescents, and found that there is a significant correlation between girls’ anxiety and depressive symptoms when they experience online harassment.

Whether in traditional or online forms, sexual harassment is a widespread and complicated problem that requires a nuanced understanding. Due to the complexity of the issue, various definitions have evolved to include a variety of inappropriate behaviors. Since a sig-

nificant portion of the population experiences online harassment to varying degrees [5], the digital age has given rise to new challenges in the field of cybersexual harassment. Young adults are particularly susceptible to these online threats [36], and the ease of perpetrating such harassment across geographical boundaries has made it a global issue [6], [7]. Understanding the nuances of both traditional and cybersexual harassment, as well as the ramifications and proactive measures that can be taken to prevent and address this type of victimization, is crucial as we navigate this dynamic environment.

2.1.3 Sexual Harassment Dataset

Due to the highly sensitive nature of the subject matter, collecting a dataset on sexual harassment proves to be an extremely difficult task [14], [17], [65]. This difficulty stems primarily from the need to respect and protect the privacy and safety of victims. Many people are hesitant to publicly share their sexual harassment experiences, which contributes to the scarcity of publicly available datasets. Furthermore, sexual harassment incidents frequently occur within specific environments, such as specific groups or organizations, limiting the diversity and scope of available data. These incidents are frequently handled confidentially or go unreported [66]–[68], making curating comprehensive public datasets difficult.

Researchers navigate the challenges of dataset collection through various approaches. Some researchers scrape datasets manually from social media like Twitter, Reddit, and Facebook, and some utilize public datasets, mainly the SafeCity dataset. [14] conducted a classification of sexual harassment stories, collected datasets from the known Sexual Harassment forums between November 2016 and December 2018, postings containing tags of sexual abuse from Reddit, and tweets about sexual harassment from Twitter. After that, they manually identified the data related to sexual harassment and got 5119 text sentences as a final dataset. [17] first discovered an open-source harassment dataset on Github, which includes 408 profane words associated with harassment, offense, or humor. Following that, they extracted tweets from Twitter that contained at least one word from the offensive keywords list, and got 3604 tweets for the classification task. [69] used two datasets on sexual harassment: Comment on Sexual Harassment (CSH), which had 212,751 comments, and

Chat Sexual Predators (CSP), which had 155,128 conversations and 2,058,781 messages. After removing text elements such as abbreviations, emoticons, digits, symbols, and varying text lengths, which made classification difficult, they got 20,000 comments from CSP and 25,000 from CSH. SafeCity platform is the biggest public online forum for reporting sexual harassment experiences. They provide 9,892 sexual harassment stories tagged as 13 forms. [70] studied sentiment analysis-based sexual harassment detection using the SafeCity dataset. [71] collected 10,622 hate speech posts from Twitter and classified them into 3 types of harassment, indirect, physical, and sexual harassment. [72] reinforced the classification model by including elements like location, time, and relationship with victims and harassers, using the SafeCity dataset as well. [73] scraped Indonesian tweets based on 11 keywords related to sexual harassment on 4 - 6 May 2022, and got 2990 tweets. [74] compared feature selection methods for sexual harassment on Facebook, using randomly selected 4000 posts from the MyPersonality dataset and 50 chat logs between volunteers and predators from the Perverted-Justice Foundation.

Some researchers create their own datasets and release them to the public. [75] provides an annotated corpus consisting of 25000 tweets and offensive word lexicons containing 5 categories: sexual, racial, appearance-related, intellectual, and political harassment content. [20] utilized the data from the top 3 most dense categories in the SafeCity dataset for sexual harassment classification, and released their dataset splits to the public.

2.1.4 Classification of Sexual Harassment Type

Studies to detect and classify the various types of sexual harassment have emerged along with the interest in sexual harassment. [17] crawled tweets including harassment content including profane words and labeled them as data implying sexual harassment and data in a non-sexually harassing. Then, they conduct binary classification with various machine learning models such as SVM, Naive Bayes, Logistic Regression, Random Forest, Gradient Boost, KNN, Adaboost, MLP, and stochastic Gradient, as well as deep learning models such as BERT, BiLSTM, CNN-BiLSTM, LSTM, CNN, and ULMFit. BERT model showed the highest accuracy of 83.56%. [76] proposed a pattern-based approach using a person iden-

tification module to detect online sexual harassment messages in social media. They used the normalization module to convert informal text that included spelling errors, slang, and contradictions to formal format and showed this module increased the classification performance. They identified patterns for sexual harassment text and achieved a 0.72 f1 score. [77] utilized a data augmentation technique called SMOTE [78] and neural networks such as CNN, LSTM, and Bi-GRU to classify harassment tweets into 4 categories (harassment, indirect harassment, physical harassment, sexual harassment). Their method showed a 0.46 f1 score, which shows the difficulty of classifying harassment data. [73] collected 2990 Indonesian tweets, used TF-IDF as a feature, and conducted binary classification using LSTM, SVM, and naive bayse. They achieved 86.54% accuracy with the SVM algorithm. [79] proposed RNN-based harassment type classification approach and back-translation method for imbalanced dataset. They classified the tweet dataset into 4 categories (harassment, indirect harassment, physical harassment, sexual harassment), and resulted 0.47 f1 score using MultiProjectedAttentionRNN. [15] proposed a model that classifies Chinese text data into four categories: general chat, uncomfortable, violated, and insulted according to the degree of harassment. They utilized a BERT-based pre-trained model and a 1-layer classifier. [70] collected data from SafeCity, a crowdsourcing platform for sexual harassment and abuse stories, and proposed the classification method that combines TF-IDF with machine learning and achieved 81% accuracy with the SGD classifier. [20] proposed an automatic classification approach working both for single-label and multi-label models to classify sexual harassment stories from SafeCity, labeled as groping, ogling, and commenting. They achieved 86.5% accuracy with the CNN-RNN model. Determining the optimal approach for the classification of sexual harassment stories depends on various factors. It is difficult to determine the best approach because the dataset, experimental conditions, and performance evaluation metrics used for each study are different. However, approaches using the neural network model, especially SVM and RNN-based models showed good results. Among the studies utilizing the English dataset, a study from [17] using the BERT model showed the highest accuracy of 83.56% in terms of accuracy, indicating the possibility that transformer-based models could be effective in identifying subtleties in sexual harassment content.

2.2 Large Language Models(LLMs)

2.2.1 Definition and Evolution of LLM

Language models are machine learning models that can comprehend and generate human languages [12], [80], [81]. Generally, language models aim to anticipate future tokens by modeling the probability of word sequences [21], [82]–[84]. They are used in Natural Language Processing(NLP) specializing in text-based data processing based on the language structure and statistical characteristics present in data. Recent studies [80], [85]–[87] demonstrated that increasing the size of language models, data sizes, and overall computation can increase model performance. This led to the advent of Large Language Models(LLMs), large-sized pre-trained language models. According to [82], LLMs are described as transformer language models, which are trained on vast amounts of text data and have hundreds of billions of parameters, such as GPT-3 [85], PaLM [87], Galactica [88], and LLaMA [89]. [21] defined LLMs as one category of artificial intelligence that has surfaced as a potent instrument for an extensive array of applications, including question-and-answer, natural language processing, and machine translation. [90] defined LLMs as models of the statistical distribution of elements in an extensive public corpus of human-generated texts. In this context, tokens including punctuation marks, words, fragments of words, or individual characters.

The first breakthrough of the neural language modeling approach was proposed by [91]. They proposed a neural language model that understands the distributed representation of words and the probability function for word sequences at the same time and demonstrated their approach improves the state-of-art approaches dramatically by experiment. Neural network based language model uses static word embeddings such as Word2vec [92] and GloVe [93], it had a fundamental problem in that it was difficult to understand the multiple meanings of a word might have in context. Researchers introduced Sequence-to-sequence learning to solve this problem [94], and it performed well in high-level language work such as machine translation [95], [96], question generation [97], image captioning [98], and speech recognition [99]. After that, deep learning based language modeling approaches appeared. Recurrent neural networks(RNN), in particular, long short term memory(LSTM) [100] was the most widely used among them [101], [102]. [103] improved traditional RNN by integrating the vec-

tor containing contextual information of the sentence. Following this, researchers have made endeavors to improve the computing complexity and network structure of RNN [104]. [105] proposed an approximate training algorithm that utilizes only a small subset of the whole vocabulary, and experiments on translation tasks. They demonstrated their method works as similar as, or outperforms the state-of-art methods on translation tasks while decreasing computing complexity. [106] proposed LightRNN, which reduces the number of vectors required for a large vocabulary by using a 2-component shared embedding, which arranges words in a table and shares row and column vectors. They experimented on language modeling tasks and achieved comparable performance to the two state-of-the-art LSTM RNN algorithms while decreasing model size and running time. However, RNN and related model and LSTM have a vanishing gradient problem, that prevents them from maintaining context in a long sentence [107], [108]. It leads to the Transformer models, which enable to use of larger data and architecture and capture longer sequences using parallel training [109].

Since the advent of the transformer model [110], researchers have conducted various studies to create more efficient and powerful language models. [111] proposed a transformer-based model that can learn 80% longer dependency than RNNs and resolves context fragmentation problem which occurs when the context is selected regardless of the meaning of a sentence, and the contextual information required to predict the following word is absent from the model. [112] proposed present Open Pre-trained Transformers, which are decoder-only pre-trained transformers, and released full models. They compared their method with existing methods and provided analysis for prompting, few/zero/one-shot, and bias/toxicity evaluation tasks.

Transformer has been such a success that almost all pre-trained models use it as the backbone. The generative pre-trained model has drawn a lot of interest lately. [113] proposed the generative pre-training(GPT) model, which utilizes the transformer model architecture and combines supervised fine-tuning with unsupervised pre-training for language understanding tasks. Following GPT, [80] proposed Bidirectional Encoder Representations from Transformers(BERT), using a pre-trained bidirectional encoder that considers both left and right context with a masked language model. They achieved new state-of-art performance on 11 NLP tasks. The pre-trained model has received great attention, and research has been

actively conducted to scale up and further improve it. Inspired by BERT, [114] proposed RoBERTa, which is the extension of BERT. They scaled up a batch size, trained the model with longer sentences, and dynamically changed the masking pattern, improving BERT on several NLP tasks. [115] proposed DistilBERT, only 40% size and 60% faster version of BERT while achieving 97% language understanding performance. They used knowledge distillation during pre-training. [116] proposed ALBERT, the model with 18x fewer parameters and 1.7x faster training speed. They used 2 parameter reduction techniques, one to divide a large vocabulary embedding matrix into two smaller matrices, and the other to prevent the parameters from increasing as the network depth increases. [117] introduced XLNet, a generalized autoregressive pretraining method. Optimizing the expected likelihood across different factorization order permutations enables bidirectional context learning and overcomes the drawbacks of BERT with its autoregressive formulation. [118] proposed pre-training framework ERINE, which takes into consideration lexical, syntactic, and semantic information in corpora during the training phase. Following this, encoder-decoder-based pre-trained models have emerged, such as T5 [119] and BART [120].

As previous studies demonstrated scaling up model parameters can improve the model performance [121], large-scale pre-trained models developed. [85] proposed the GPT-3, autoregressive language model trained with 175 billion parameters. They experimented with GPT-3 on one, zero, and few-shot settings on various NLP tasks such as language modeling and question answering, and demonstrated the strength of GPT-3 for generating high-quality answers without fine-tuning or gradient updates. Similarly, large models with billions of parameters such as PANGU [122], GShard [123], Switch-transformers [124] developed. More recently, MetaAI proposed LLaMA [89], containing 7B to 65B parameters and trained with trillions of tokens. Google Research [125] proposed Pathways Language Model (PaLM), a densely activated Transformer model including 540-billion parameters.

The majority of language models were pre-trained with English, and some researchers pre-trained models on different language corpus and improved performance on specific languages. [126] used French corpus and [127] used Korean corpus for pre-training.

2.2.2 Applications of LLM

LLMs can be widely used and are rapidly evolving in various applications. LLMs are beneficial for several natural language processing applications including natural language processing tasks such as text classification [128]–[130], text and content generation [131]–[133], information retrieval [134], [135], chatbot systems such as educational systems [136] and virtual assistants [137], machine translation tasks [138]–[140], text mining [141]–[144], data summarization [145]–[148], and speech processing [149]–[151]. Due to LLMs’ wide range of applicability, researchers actively conducting studies that employ LLMs to address social problems.

[152] explored the advantages and disadvantages of LLMs as well as their potential to increase the effectiveness and efficiency of clinical, educational, and research work in medicine. [128]–[130] proposed text classification using a large language model along with various techniques such as transfer learning and fine-tuning. [153] proposed Plug and Play language model, which combines pre-trained language model with attribute classifiers and enables attribute controlling for text generation. [154] use the word representations as the model training parameter, and improve the limitation of high cost in large search space of existing lexically-constrained text generation approaches. [155] proposed sentiment controllable dialogue generation model, which generate the conversation text contains specified emotion implicitly or explicitly. [156] proposed the storytelling framework that user can give multiple topics to the model and model generates the story based on given topics. [133] utilized BART and T5 model for graph-to-text generation task, and achieved new state-of-the-art results. [157] introduced Seq2Seq constrained keyword-based text generation technique. [158] explored the few-shot data augmentation technique for information retrieval tasks. In the medical field, LLM has also been broadly utilized in chatbots; studies have been done comparing the responses of chatbots based on LLM with those of real doctors. [159] compared consent forms of 6 surgical procedures generated by LLM based chatbots and surgeons, and found that chatbot-generated consent forms are simpler, and get better scores for completeness and accuracy. [160] compared LLM based chatbot’s response and Ophthalmologist’s response about patient eye care questions, and patients could only distinguish

between chatbots and real people’s answers with 61.3% accuracy. Furthermore, LLM-based chatbots are also used to give health advice on heart attacks [161] and are used in the field of education [136]. [138] studied the effectiveness of large language models in machine translation, and proposed a novel smoothing method for training large data sets. Additionally, LLM is actively utilized in the field of machine translation. [140] examine the capabilities of LLMs such as GPT-3.5, GPT-4, and BLOOM for real-time machine translation tasks, and suggest a machine translation system that is effective for languages with fewer resources. LLM is also widely used in text mining and data summarization tasks in various domains. For instance, [142] proposed SMedBERT for medical text mining, [143] proposed BioBERT for biomedical text mining, [144] proposed Finbert for financial text mining, and [141] proposed MatSciBERT for materials domain text mining. Similarly, [147] proposed the data summarization model BioBERTSum for the biomedical field using a sentence position embedding mechanism. [162] introduced a large-scale multi-document summarization dataset and model for news articles. [163], [164] focused on long-document summarization task. Moreover, LLMs are also employed in tasks involving speech recognition. [149] compared several language model integration methods for speech recognition, both on medium-sized and large-sized datasets. [151] proposed automatic speech recognition system and outperform monolingual baselines by adding a conformer encoder to LLaMA model. In addition to these examples, LLMs are actively applied and vastly evolving in a wide range of fields, and researchers developing new models and applications at a very rapid pace.

There are multiple studies using LLMs for addressing social issues. [165] explores and evaluates LLMs for real-world security challenge scenarios including cryptography and reverse engineering. [17] use LLM for online sexual harassment classification. [16] proposed ‘Llama guard’, an LLM-based safety tool targeting human-AI interaction. This tool is useful for classifying specific safety risks arising from LLM prompts and detects conversations that violate safety risk guidelines such as violence& hate, sexual content, guns & illegal weapons, and self harm. [166] and [167] proposed hate speech detection techniques using GPT-based LLM and showed that LLM-based detection method surpassed existing approaches. [168] proposed fine-tuned BERT to detect cyberbullying and demonstrated that BERT outper-

formed CNN, LSTM, or BiLSTM models. [169] explores LLMs’ ability to detect implicit hate speech and their ability to express confidence in their responses.

2.3 Out-of-Vocabulary(OOV) Word

2.3.1 Definition

Word embeddings are widely used as a feature for neural networks in various NLP tasks, but Out-of-Vocabulary(OOV) words degrade the model performance due to the information loss [170]. OOV words are words that are not in the model’s training set. OOV words consist of new words that come from various sources such as scientific and engineering terms, new terms from social life, political terms, and foreign words [171]. They also contain typos and slang words [172].

2.3.2 OOV Word Handling

The OOV word problem has been approached in the literature in several ways. Several works use morphemes to generate the embeddings for OOV words. [173] utilized morpheme vectors to compute word embeddings. [174] proposed word embeddings that combine morphological and distributional information. [175] propose the morphological structure based OOV embedding generation method that is trained with the function of spelling distribution of words using Bi-LSTM architecture. Their method is useful for low-resource languages. But morpheme-based approaches struggle with foreign language words and names. Other approaches use character-level language models or embeddings, which work directly with word characters rather than pre-established word tokens. [176] suggested the open-vocabulary word embedding model that generates the word embedding using RNN and character-level embeddings. [177] proposed the language model using character-level inputs employing CNN and RNN, and outperformed morpheme-based models with fewer parameters. [178] proposed the embedding model that utilizes character n-gram count vector and single layer transformation. [179] proposed a several subword modeling approach to get word representations from characters and morphemes of a word. Their method using morphemes showed strong performance on machine translation tasks for morphologically rich languages. [180] introduced

the FastText model that can encode rare words. They use the sum of character n-grams for word representation and use subword units to model morphology.

Furthermore, some researchers explored techniques that incorporate contextual information and deep learning methods. [175] suggested the method that considers both the morphology and context of the word to predict embeddings for OOV words. They employed recurrent network and attention mechanisms to capture the right and the left contextual information. [181] proposed HiCE, the method to construct the OOV embeddings using attention-based architecture on a few-shot learning setting. [182] extended the work of [175] by considering the embeddings of surrounding words as well as the characters of the word. They proposed a network called 'Comick' that predicts the embeddings of OOV words using morphology, contextual information, and an attention mechanism. [183] proposed a method to adapt unsupervised word embeddings for noisy and small datasets using a neural network with one single hidden layer. They estimate embeddings into a low dimensional sub-space that enables the embeddings to fit the complexity of the target task. [184] proposed a POS tagging system using a deep neural network. They predict the word embedding for OOV words using semantic and morphological information using a pre-trained model learned by FastText embeddings and Bi-LSTM layer architecture. [185] proposed the word embedding model for Korean Characters(Hangeul), using Convolutional Neural Network(CNN) architecture with an attention mechanism. Their model demonstrated the robustness under a high-noise level environment.

Some works focused on embedding generation for specific tasks, such as Part-of-Speech (POS) tagging or machine translation tasks. [176] proposed the sequence tagging method to build the word embeddings by compositing the characters using Bi-LSTM. [186] proposed the model for rare words, utilizing the word log frequency and auxiliary loss. [187] proposed the neural machine translation model that is robust to various kinds of noise using adversarial training and structure-invariant representation. [172] proposed the contrastive learning framework for building a robust embedding model for OOV words. They use a mixture of characters and sub-words as input.

Several different methods have been suggested for handling OOV words in addition to the ones already described. [188] proposed the method to calculate the average embedding

of the context for unknown words. [24] proposed a multi-level OOV handling method that resembles the human brain’s inference system. Their approach is composed of 3 phases, analogy, decoding, and prediction, combining multiple strategies to resolve limitations of previous approaches that focus on specific types of OOV words. In the text classification task, their approaches demonstrated competitive performance in the majority of experiments on noisy and short text datasets. [189] proposed the Misspelling Oblivious (word) Embeddings (MOE), the method to generate embeddings that are robust to typo and misspelling by learning from real typo data collected from the web. [190] propose a model that is robust to typos and misspellings, using the method to encode words as character sequences. They divide the target word into 3 parts, Beginning, Middle, and End vectors to generate the embeddings. They experimented with 3 languages(English, Russian, and Turkish) and 3 tasks, paraphrase generation, sentiment analysis, and textual entailment identification. Their method is more robust to typos than word2vec and FastText models especially when noise level is high. [191] proposed the variant of BERT, CharacterBERT that modifies the tokenizer process to subword level and improves performance on the the less general domains.

2.4 Summary

This chapter provided a review of the literature relevant to sexual harassment, large language models, and OOV handling. The next chapter provides the framework and methodology to be used in the research project.

3. METHODOLOGY

This chapter provides the framework and methodology to be used in the research study.

3.1 Study Design

This study aims to conduct an analysis of LLM’s performance of sexual harassment story classification tasks before and after replacing Out-of-Vocabulary words. The research seeks to identify the subtle impacts on the interpretability, sensitivity, and general performance of the ensuing sexual harassment narrative categorization task of LLMs. The author replaces OOV words, conducts text classification with the BERTForSequenceClassification model on the dataset before and after replacement, and compares their performance before and after replacement.

3.2 Dataset

3.2.1 Dataset Description

The author uses the SafeCity dataset karlekar2018safecity to train our model for the classification. SafeCity platform is the biggest public online forum for reporting sexual harassment experiences, and sexual harassment stories are submitted and tagged as 13 forms by the forum users. The stories were mainly collected in India, and other countries such as Kenya, Nepal, and Malaysia. The author uses the dataset provided by [20], which contains 9,892 stories from the top 3 categories, commenting, ogling, and groping from the SafeCity dataset. The dataset is composed of stories categorized into different types of sexual harassment. Verbal sexual harassment-related behaviors were classified as a "Commenting" class, visual sexual harassment-related behaviors were classified as an "Ogling" class, and physical sexual harassment-related behaviors were classified as a "Groping" class. Examples of "Commenting" class could include making sexually explicit comments or jokes, and bullying or humiliating someone using sexually suggestive language. Behaviors such as following someone with one’s gaze focusing on their body, leering at someone with obvious sexual intent, and whistling while staring at someone’s body could be examples of "Ogling" class.

Table 3.1. Example of SafeCity dataset

Description	Commenting	Ogling	Groping
a girl was teased by a boy and he was showing some facial expression to the girl.	0	1	0
Me and my friends were coming back from our internship and going towards the Dwarka sector 10 metro station. There is a wine and beer shop where a lot of men hang around and pass comments and behave indecently.	1	1	0
This happened to me in the last 30 days during the day. I took a shared auto rickshaw to reach the malviya nagar market and a man touched me inappropriately.	0	0	1

In the "Groping" class, behaviors indicating grabbing someone's body, suggestively touching someone's body, and pinching or squeezing someone's body parts in a sexual manner without permission can be included. Specifically, 39.3% of the dataset pertains to "Commenting," 21.4% to "Ogling," and 30.1% to "Groping." This dataset contains both single-label and multi-label datasets, consisting of 8,191 training samples and 1,701 test samples for each category. Single-label dataset is for binary classification, and it consists of 2 columns, description and category. If the story in the description column contains sexual harassment it is labeled as 1, and if not it is labeled as 0. Multi-label dataset is for multi-class classification and it contains 4 columns, including description, commenting, ogling, and groping. The corresponding type of sexual harassment is labeled 1 if it is included in the story and 0 if not, and some stories include one or more sexual harassment components. The author use a multi-label dataset for this study. Table 3.1 is an example of a multi-label dataset.

Table 3.2. Dataset Distribution

Class	Label	# of Instances
Commenting	0	2,013 (40.5%)
Ogling	1	668 (13.4%)
Groping	2	2,321 (46.4%)
Total		5,002

3.2.2 Dataset Preprocessing

The author utilized the multi-label dataset, excluding entries associated with two or more categories, thus focusing solely on instances belonging to a single category. The dataset had 5,002 stories that were unique to one category. The author re-labeled the dataset as commenting, ogling, and groping. There are 2,013 instances (40.5%) for Commenting, 668 instances (13.4%) for Ogling, and 2: 2,321 instances (46.4%) for Groping. Among the entire dataset, 51.2% contain at least one out-of-vocabulary (OOV) word. The author lower-cased and lemmatized all data, and tokenized the data with BERT Tokenizer. The final dataset distribution is described in Table 3.2.

3.3 Models

The author used DistilBERT for the replacement of OOV words, and BERT for the classification.

In the process of replacing out-of-vocabulary (OOV) words with in-vocabulary words from the dataset, a step involves masking the OOV words and using a language model to predict the masked words. Here, DistilBERT was employed for this task. The detailed procedure will be discussed in the section 3.4.

The author used a BertForSequenceClassification model based on *bert-base-uncased* from HuggingFace with the SafeCity dataset for the classification. *bert-based-uncased* model includes 12 layers, 12 attention heads, and 768 hidden sizes with 110M parameters. It utilizes the BERT model, which comprises several layers, including embeddings, encoder, and pooler layers. The embeddings layer handles token, position, and token type embeddings. The en-

coder layer consists of multiple BertLayers, each containing attention, intermediate, and output sublayers. Finally, the pooler layer aggregates token representations for classification. The model also includes dropout layers for regularization. The structure of the model the author used in this study is proposed in Figure 3.1.

There are several reasons for the selection of BERT based model in this study. According to the study conducted by [17], BERT showed the best performance on the sexual harassment story classification task when compared to other machine learning algorithms and deep learning algorithms. In addition, The BERT-based models are the most widely used for text classification tasks in the HuggingFace platform. These models leverage the BERT architecture and undergo fine-tuning tailored to their specific objectives. This approach has been prevalent due to its effectiveness in achieving state-of-the-art performance across various natural language processing tasks. The most of previous studies on sexual harassment story classification use CNN, RNN, or Bi-LSTM-based classification models. Since this study focuses on the LLM, the author utilized the BERT model for the classification task. [192] proposed the strategy to select a suitable approach to using LLM in text classification. In our study, document length is short, a number of documents is more than 1,000, time allocation is not limited, and the budget is limited. In this case, they recommend annotating a large sample and fine-tuning a smaller model. The author already has a fully annotated dataset, and BERT is smaller than GPT-3 and widely used in text classification tasks, which makes BERT an appropriate selection. In addition, a previous study by [192] showed that fine-tuned BERT performed almost as well as the GPT-3 model, which is larger, in most NLP tasks.

BERT (Bidirectional Encoder Representations from Transformers) devlin2018bert BERT is a language model that has been pre-trained using unlabeled text data. It was proposed by Google and has significantly advanced natural language processing. It is built on the Transformer architecture, and featured with its capacity to encode words bidirectionally to take into account context from both directions within a sentence. This allows for a richer comprehension of the context in actual conversations or papers.

DistilBERT (Distilled Bidirectional Encoder Representations from Transformers) sanh2019distilbert DistilBERT is a lightweight version of BERT that was proposed

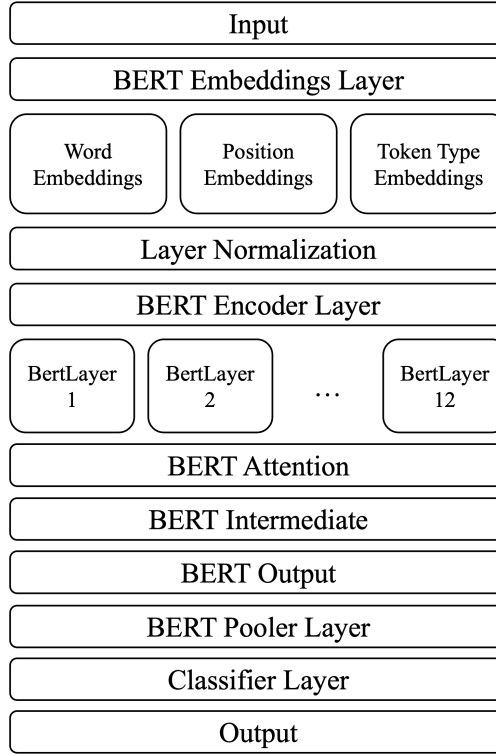


Figure 3.1. BertForSequenceClassification Model Structure

by HuggingFace. DistilBERT is based on the BERT architecture but focused on reducing the number of parameters to lower the computational cost and footprint of the model. Even with a large reduction in parameters over BERT, DistilBERT preserves the important aspect of BERT: bidirectional context encoding. This indicates that it can effectively encode words and retain performance across a range of natural language processing tasks even with a reduced model size. DistilBERT is extensively used in settings with limited resources and in applications that call for effective NLP models.

3.4 OOV Replacement

Since BERT uses the WordPiece algorithm to create vocabulary `schuster2012japanese`, the author uses the WordPiece algorithm to detect OOV words. The author obtained a list of in-vocabulary words using BERT Tokenizer, and for each sentence in the dataset,

the author selected OOV words that were not included in the vocabulary list. Based on [24], the author adopt a multi-level approach that resembles the human thought process maluf2013alfabetizaccao to replace OOV words. In comparison with various existing OOV handling approaches such as fastText [180], HiCE [181], and Comick [182], experiments have demonstrated that their multi-level OOV handling methods exhibit generally satisfactory results across datasets encompassing noisy text and diverse contexts. It is composed of 3 steps, analogy, decoding, and prediction.

In the Analogy step, the author replaced OOV words with synonyms found in thesaurus dictionaries. The author modified the existing approach by incorporating a prediction model to select the most suitable word among the initial synonyms obtained in the first step. This addition was made considering that the dataset belongs to a narrow domain and to prepare for instances where suitable in-vocabulary synonyms for replacement words cannot be found in the dictionary. Next, in the Decoding step, the author replaced OOV words with words having similar morphological structures. Additionally, in cases where an appropriate in-vocabulary word could not be found even after the decoding step, the author utilized the average embedding of in-vocabulary words among the synonyms obtained in the first step.

In the first step, the author used a dictionary and masked language model to search for synonyms of OOV words. The author used the [193] API and retrieved the top 10 synonyms for each OOV word. Datamuse API returns related words with the original OOV word even if it is a foreign language or the name of a person. For example, the term 'matatu' means privately owned minibusses used as shared taxis in Kenya. Datamuse API returns words like 'local taxi', 'minibus', and 'share taxi' which are relevant to and fully reflect the original meaning of the OOV word. For the same word, there were no synonyms found in the WordNet dictionary. Also, datamuse API finds the synonym based on the context information of the word. For instance, for the word 'Pashupatinath' which means Hindu temple in Nepal, WordNet returned nothing but datamuse returned some words related to the place of the temple and Hindu religion. The author selected the top 10 words because based on the empirical observations by [24], datamuse API generally turns return noise after 10 words. Among 10 candidates obtained from datamuse API, the author select words that are in-vocabulary words.

After extracting candidate replacement words for OOV words from datamuse and masked language prediction model, the author found common words among the words obtained from each method. If no match was found, the author checked whether the word was a typo or a spelling error using the spelling checking library [194]. Symspellpy is a spellchecking library using Levenstein distance, and it enhanced the searching speed by precomputing possible spelling errors for the dictionary. The author selected only the words with a Levenshtein distance of 2 or less from the original word and chose only those that are in-vocabulary words.

If the author couldn't find replacement words for OOV words in the previous step, the author used the average embedding values of the in-vocabulary words among the synonyms obtained from the datamuse API. In case there's no in-vocabulary word in datamuse API synonyms list, the author used the average embedding values of the in-vocabulary words predicted from the DistilBERT masked language prediction model. If there are multiple words in the list, the author selects the word with the most similar embedding value with the average embedding value. The author employed cosine similarity when calculating the similarity of embedding value. For the embedding model, the author chose the BERT embedding method. A study conducted by [195] compared the text classification performance using various word embedding methods such as FastText, word2vec, ELMo, GloVe, and BERT with various classification datasets. The BERT model outperformed other embedding methods in text classification tasks on single-label datasets. In cases where the entire sentence or specific words are in a different language, those sentences are manually translated to English sentences with the same meaning. Our replacement approach is described in Figure 3.2.

3.5 Classification

The aim of this study is to assess and compare the performance of Large Language Models (LLMs) for the task of sexual harassment story classification before and after the replacement of OOV words. The author evaluated the model's classification performance in this analysis. The investigation will focus on analyzing the data that the model correctly and incorrectly classifies. The purpose of this analysis is to identify differences in the features and patterns

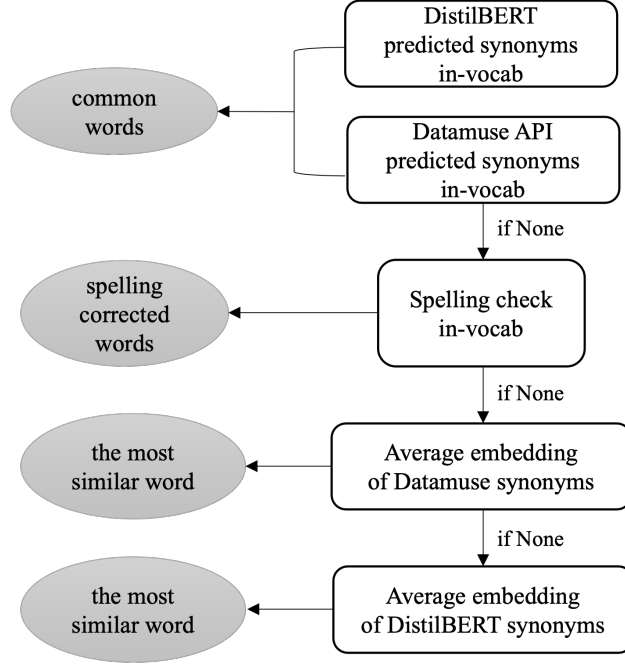


Figure 3.2. Proposed OOV word replacement flow

of correctly classified cases versus incorrectly labeled cases, illuminating the advantages and disadvantages of the large language model for correctly classifying sexual harassment stories. A thorough investigation into the incorrectly classified cases will also be conducted to identify the particular difficulties or subtleties that cause mistakes in the classification procedure. Comparing the classification performance before and after OOV replacement allows for a better understanding of the large language model’s domain understanding and generalization capabilities. This provides insights into how effectively the model operates within specific domains.

3.5.1 Dataset splitting and Cross-validation

The author performs a 3-class classification on a fine-tuned bert-based-uncased model with the SafeCity dataset and compares the performance of each model. The author uses 80% (3,627 instances) of the dataset for training and 20% (1,375 instances) for testing. The author use 20% of the training dataset for the validation set. The author divides the dataset

into 5 folds and performs cross-validation for reliable results. The author takes the average value of cross-validation as the final result. For each model, the author conducts classification twice, using the original dataset and the OOV-replaced dataset.

3.5.2 Model Training

The author used the BertForSequenceClassification model with 1 classification layer. For training, the author utilized the AdamW optimizer with a learning rate of 2e-6. Additionally, a linear scheduler with warmup was employed to adjust the learning rate during training. The number of epochs was set to 10. These settings were chosen to balance model performance and training efficiency. Deliberate parameter optimization and fine-tuning were omitted, as the primary objective of this study was not to engineer a high-performance model, but rather to compare the model's performance pre- and post-replacement of OOV words. The author trained the model on the original dataset and replaced the dataset with the same train and test set.

3.6 Result Analysis

3.6.1 OOV Feature Analysis

The author analyzed examples where the model made correct and wrong classifications to study the types of out-of-vocabulary (OOV) words in each scenario. The author used TF-IDF (Term Frequency-Inverse Document Frequency) to identify if an out-of-vocabulary word was a term related to the sexual harassment domain. TF-IDF is a statistical metric that evaluates the importance of a word in a document compared to a set of documents by multiplying the term frequency by the inverse document frequency. The author calculated the TF-IDF values for each out-of-vocabulary (OOV) word in the document using the complete Safecity dataset as the reference text. The author labeled out-of-vocabulary (OOV) words as relevant to the dataset if they had TF-IDF values of 0.64 or higher in our analysis. This threshold was chosen because, upon investigating the TF-IDF values calculated for each word, the author found that values below 0.64 were indicative of words unrelated to the sexual harassment domain.

3.6.2 SHAP Analysis

The author uses the SHapley Additive exPlanations (SHAP) analysis proposed by [196]. SHAP is the game theory based method to explain the impact of each input feature on the model's prediction by comparing predictions with and without each feature. It employs simplified explanation models to provide insights into complex machine learning models. To analyze which words had the greatest impact on the model's predictions for correctly and incorrectly classified instances in each dataset, the author employed the SHAP method. For the language model, the SHAP baseline value is the model output when the entire sentence is masked. SHAP values explain how unmasking each word impacts the model output by changing from the baseline value to the final prediction value in an additive manner. Positive SHAP values indicate that a feature positively contributes to increasing the model's output or likelihood of a certain class prediction. Conversely, negative SHAP values indicate that a feature negatively contributes to the model's output or decreases the likelihood of a certain class prediction. For example, in Figure 3.3, the saliency plot shows how each input feature affects the model's prediction for each class.

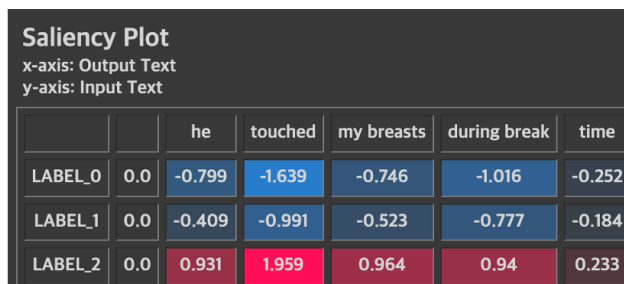


Figure 3.3. SHAP Analysis Saliency Plot

The example sentence belongs to the 'LABEL 2' class. Figure 3.4 represents a graph where the SHAP values indicate how much each word influenced the model's classification of this sentence into the 'LABEL 2' class. Higher SHAP values suggest that the corresponding word had a significant impact on the model's decision to classify this sentence into that particular class.

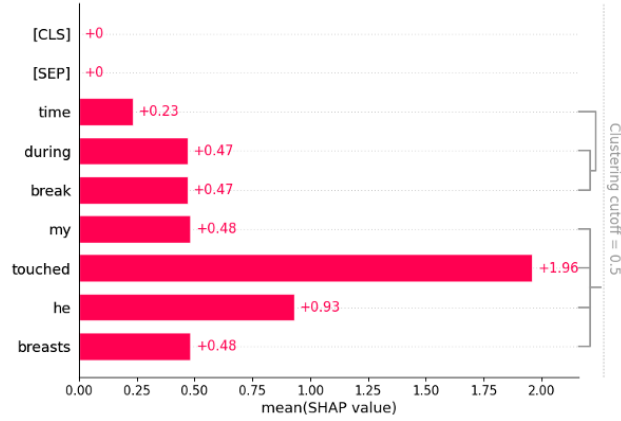


Figure 3.4. SHAP Analysis Bar Graph

In this study, the author used the SHAP value to figure out the word that has the highest impact on the model's prediction and conducted the comparative analysis on the model trained on the original dataset and the model trained on the dataset after OOV replacement.

3.7 Summary

This chapter provided the framework and methodology to be used in the research study. The next chapter provides experimental results.

4. RESULTS

This chapter presents the experimental results, focusing on assessing the model’s performance before and after replacing out-of-vocabulary (OOV) words. The author delves into three scenarios: instances where the model classified data correctly before but incorrectly after OOV replacement, vice versa, and where classification results remained consistent. Additionally, the author examines the characteristics of OOV words in original data where the model made correct and incorrect classifications. Employing the SHAP method, the author investigates the impact of individual words on the model’s predictions for both correctly and incorrectly classified instances across datasets. Through this comprehensive analysis, the author aims to understand the sensitivity of the model to OOV words and their behavior in a narrow domain with and without such words.

4.1 Evaluation Metrics

The author used the following metrics to measure the performance of each model: accuracy, precision, recall, and F1 score. In a classification task, the precision is the number of true positives divided by the total number of elements classified as positive class by the model. Recall is the number of true positives divided by the total number of elements that actually belong to the positive class. True Positive here means the number of items correctly labeled as belonging to the positive class. Accuracy is the measurement of error. This indicates how close a given result is to its true value. Accuracy in multi-class classification is defined as the number of correct classifications divided by the number of all classifications. F1 score is the harmonic mean of the precision and recall.

The model trained on the original dataset achieved an accuracy of 80%, with precision, recall, and F1-score values varying across classes. Class 0 (Commenting) exhibited a precision of 79% and a recall of 84%, while Class 1 (Ogling) had a precision of 81% and a recall of 53%. Class 2 (Groping) showed the highest precision of 82% and the highest recall of 85%. The model trained on the replaced dataset, on the other hand, achieved a slightly lower accuracy of 79%. Despite similar precision values, the OOV replaced model displayed slightly lower recall values across Class 0 and Class 1, and showed slightly higher recall value in Class 2

Table 4.1. Model Performance on Original Dataset

Class	Precision	Recall	F1-Score
Commenting	0.79	0.84	0.81
Ogling	0.81	0.53	0.62
Groping	0.82	0.85	0.75
Accuracy	0.80		

Table 4.2. Model Performance on Replaced Dataset

Class	Precision	Recall	F1-Score
Commenting	0.81	0.82	0.82
Ogling	0.81	0.43	0.54
Groping	0.81	0.86	0.81
Accuracy	0.79		

Table 4.3. The most influential words for model prediction (Original dataset)

Label	Correct	Incorrect
Commenting	commenting, commented, comments, boys, harassment	its, staring
Ogling	staring, og, ling, stared, was	guys, touched, at, for, and
Groping	touched, touch, and, raped, touching	boys, pictures, cal, a, teasing

compared to the original model. The detailed model performance on the original dataset and the dataset after OOV replacement is shown in Table 4.1 and Table 4.2.

4.2 SHAP Analysis

The author aimed to identify the most influential words that the model relied upon when classifying data into each class using SHAP. The author analyzed specific words that played a crucial role in the model’s classification decisions for each class. The author selected the most frequently occurring words among those identified to have the highest positive SHAP values.

Starting with the model trained on the original dataset, for the "Commenting" class, when the model correctly classified data belonging to this class, the frequently mentioned words that had the most significant impact on the model’s decision were 'commenting,' 'commented,' 'comments,' 'boys,' and 'harassment.' On the other hand, in cases where

Table 4.4. The most influential words for model prediction (Replaced dataset)

Label	Correct	Incorrect
Commenting	commenting, commented, comments, and, comment	staring
Ogling	staring, gazing, and, was	a, my, following, touched, at
Groping	touched, touch, and, touching, raped	a, and, was

data belonging to this class were misclassified, 'its' and 'staring' were selected as the words that most frequently influenced the model's decision. For the "Ogling" class, when the model classified data correctly, 'staring,' 'og,' 'ling,' 'stared,' and 'was' were identified as the most influential words. However, when data was misclassified, 'guys,' 'touched,' 'at,' 'for,' and 'and' were the words most frequently influencing the model's decision. In the groping class, when the model correctly classified data, 'touched,' 'touch,' 'and,' 'raped,' and 'touching' were frequently mentioned as the most influential words. Conversely, when the model misclassified data in this class, 'boys,' 'pictures,' 'cal,' 'a,' and 'teasing' were the words most frequently influencing the model's decision.

For the dataset after the OOV replacement, notable discrepancies in the classification results were observed. Specifically, concerning data within the "Commenting" class, instances correctly classified featured 'commenting,' 'commented,' 'comments,' 'and,' and 'comment' as the most influential words. Conversely, for misclassified data, 'staring' was frequently cited as having a significant influence. Regarding the "Ogling" class, correct classifications were characterized by 'staring,' 'gazing,' 'and,' and 'was' emerging as the predominant influential words. Conversely, instances of misclassification within this class were associated with 'a,' 'my,' 'following,' 'touched,' and 'at' as the most influential terms. Within the "Groping" class, instances correctly classified were marked by 'touched,' 'touch,' 'and,' 'touching,' and 'raped' as the frequently mentioned influential words. Conversely, instances of misclassification within this class were linked with 'a,' 'and,' and 'was' as the most influential terms. Word lists are described in the Table 4.3 and Table 4.4.

4.3 Comparison of Model Prediction on the Original Dataset and Replaced Dataset

Together with these statistical performance indicators, the author analyzed data features in case the model classifies correctly and incorrectly before and after replacing OOV words to get a deeper understanding of the results.

The author analyzed the data feature using SHAP analysis in three cases: 1) The data is classified correctly before replacement but incorrectly after replacement (C-I), 2) The data is classified incorrectly before but correctly after replacement (I-C), and 3) The data is classified both correctly (C-C) or incorrectly (I-I) before and after replacement. The author found that out of a total of 1,375 test instances, 862 instances were correctly classified by both models, while 370 instances were misclassified by both. Also, 84 instances that were correctly classified in the original dataset were incorrectly classified in the replaced dataset, and 59 instances that were incorrectly classified in the original dataset were correctly classified in the replaced dataset. Figure 4.1 describes the number of instances in each case.

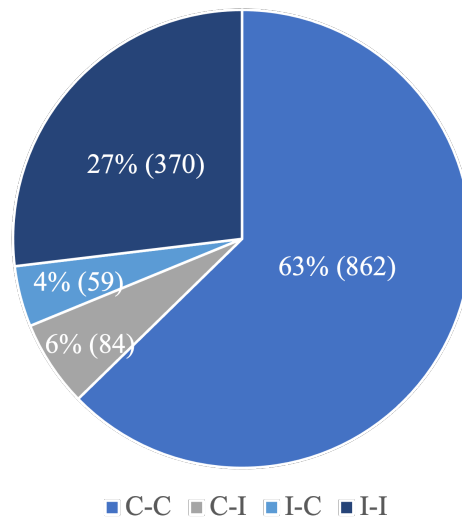


Figure 4.1. Analysis of Correct and Incorrect Classifications Before and After Replacement

4.3.1 Same Results Before and After OOV Replacement (C-C, I-I)

For a total of 1,232 instances, the model exhibited consistent results before and after replacement. Out of these, 862 instances were correctly classified in both the pre- and post-replacement datasets, while 370 instances were consistently misclassified in both datasets.

The author utilizes SHAP values to identify the word that has the most influence on the model's prediction. In the "Commenting" class, for instance, the same keyword was identified as the most influential for 177 instances before and after replacement, while different keywords were selected for 144 instances. Similarly, in the "Ogling" class, 24 instances had the same influential word, while different words were identified for 29 instances. The "Groping" class exhibited a similar pattern, with 193 instances having the same influential word before and after replacement, while 295 instances had different influential words. Out of the total 862 instances correctly classified before and after OOV replacement, 54% (468 instances) had different words identified as most influential. Although there were more instances in the "Commenting" class where the most influential word remained the same before and after replacement, the other two classes saw a higher number of instances where the influential word changed.

The analysis of instances incorrectly classified both before and after replacing OOV words yields the following findings. In cases where the same word was identified as most influential both before and after replacement, there were 4 instances in the "Commenting" class, 24 instances in the "Ogling" class, and 99 instances in the "Groping" class. Conversely, instances where different words were identified as most influential before and after replacement were 17 instances in the "Commenting" class, 44 instances in the "Ogling" class, and 182 instances in the "Groping" class. Out of a total of 370 instances, only 28% (127 instances) had the same word identified as the most influential before and after replacement.

In Figure ??, the most frequently selected common word when the model chose the same word as the most influential before and after OOV word replacement is listed. When the model correctly classified the class of the data, the most influential words were primarily words that were relevant to the dataset and related to each class across all three classes. In the "Commenting" class, 'Commenting', 'comments', and 'commented' were selected as

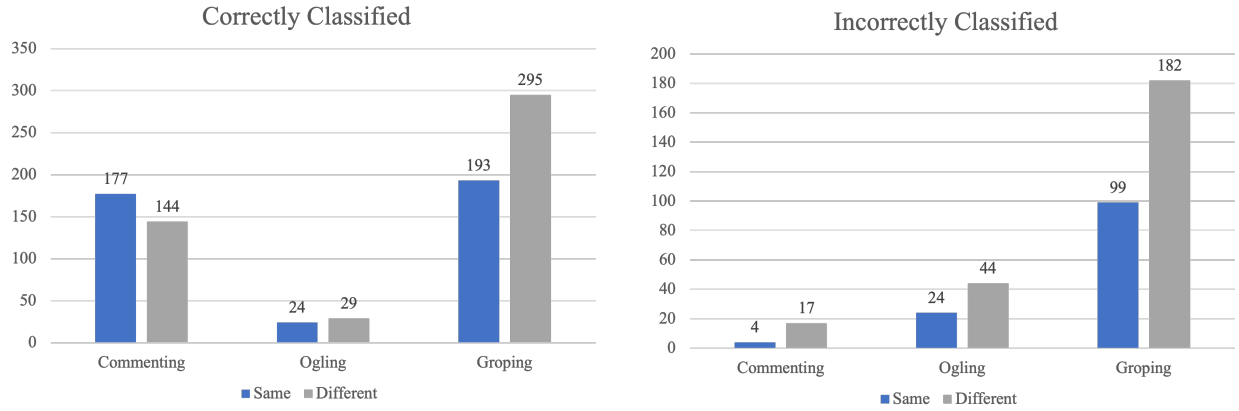


Figure 4.2. Comparison of Consistency in Outcome Between Most Influential Words Selected Before and After OOV Replacement

Table 4.5. Comparison of Predicted Classes Before and After OOV Replacement (I-I)

Label	Same	Different
Commenting	20	1
Ogling	60	8
Groping	267	14

the most influential words, while in the "Ogling" class, 'staring', and in the "Groping" class, 'touch', 'touched', and 'touching' were identified. This indicates that in both the pre- and post-replacement models when correctly classifying the classes, the most influential words selected were the words that closely related to the meaning of the name of each class. Conversely, when the model incorrectly classified the class of the data, it selected words that were relevant to the dataset but unrelated to the respective class or entirely unrelated to the domain as the most influential words. For example, in the "Commenting" class, words like 'staring', 'and', and 'looks' were present, in the "Ogling" class, 'bus', 'at', 'stalking', and in the "Groping" class, 'and', 'whistling', 'harassment' were identified.

Table 4.5 represents the number of instances predicted for each class that are the same and different when the model predicts both incorrectly before and after the OOV replacement, and Figure 4.3 describes the number of predictions in each class. In the "Commenting" class, both models predicted 20 cases as the same class before and after replacing OOV words,

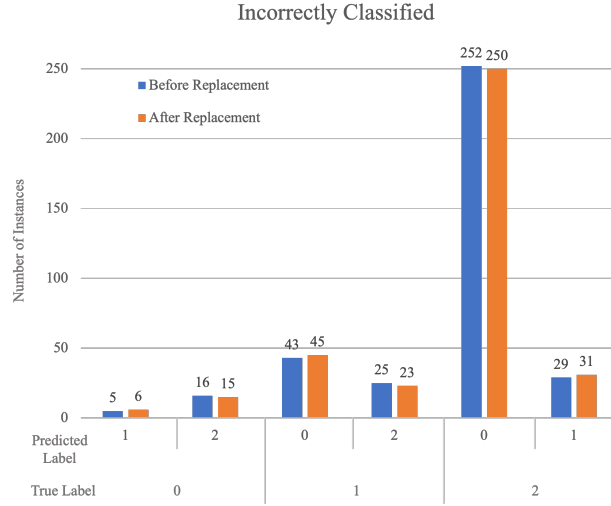


Figure 4.3. Comparison of Predicted Classes Before and After OOV Replacement (I-I)

with only 1 instance being predicted differently. Among them, for the model before the replacement, 5 of them were predicted as the "Ogling" class and 16 were predicted as the "Groping" class. For the model after the replacement, 6 instances were classified as the "Ogling" class, and 15 were classified as "Groping" class. For the "Ogling" class, there are 60 instances classified to the same class before and after OOV replacement, and 8 instances classified to the different classes. The model before replacement classified 43 instances to the "Commenting" class and 25 to the "Groping" class. On the other hand, the model after replacement classified 45 instances to the "Commenting" class and 23 instances to the "Groping" class. The majority of instances for the I-I case fall into the "Groping" class. In the "Groping" class, 267 instances are classified to the same class by both models, and 14 are classified to a different class. Before the replacement, 252 instances are classified as the "Commenting" class and 29 are classified as the "Ogling" class. After OOV replacement, 250 instances were classified as the "Commenting" class, and 31 were classified as the "Ogling" class.

Table 4.6. Top Common Keywords Before and After OOV Replacement When Correctly Classified

Class	Keywords
Commenting	'commenting', 'comments', 'commented', 'harassment', 'and', 'bad', 'on', 'were', 'the'
Ogling	'staring', 'and', 'was', 'guy', 'pm', 'any', 'college', 'showed', 'happens', 'disturbed'
Groping	'touch', 'touched', 'touching', 'and', 'my', 'me', 'her', 'tried', 'guy', 'was'

Table 4.7. Top Common Keywords Before and After OOV Replacement When Incorrectly Classified

Class	Keywords
Commenting	'staring', 'and', 'looks', 'touches'
Ogling	'bus', 'at', 'stalking', 'harassment', 'travelling', 'side', 'it', 'avail', 'pictures', 'street'
Groping	'and', 'whistling', 'harassment', 'were', 'bad', 'teasing', 'happened', 'pictures', 'stalking', 'a'

4.3.2 Different results before and after replacement (C-I, I-C)

Out of a total of 1,232 instances, 84 instances that were correctly classified in the original dataset were incorrectly classified in the replaced dataset. Among them, 15 instances belong to the "Commenting" class, 6 instances belong to the "Ogling" class, and 63 instances belong to the "Groping" class. Considering the dataset distribution shown in Table 3.2, it can be observed that the number of instances in the "Commenting" and "Groping" classes does not differ significantly. However, in the replaced dataset, the number of instances incorrectly classified is much higher in the "Groping" class. Therefore, when comparing the original dataset with the replaced dataset, it is evident that the highest increase in errors occurred in the "Groping" class.

On the other hand, 59 instances that were incorrectly classified in the original dataset were correctly classified in the replaced dataset. Among them, 12 instances belong to the "Commenting" class, 3 belong to the "Ogling" class, and 44 belong to the "Groping" class. Similarly, the number of instances in the "Groping" class is relatively high compared to the total number of instances.

Table 4.8 describes the number of instances the model select the same influential word before and after OOV replacement. In the case of C-C, out of a total of 84 instances, the model selects the same word as the most influential word before and after replacement in 14

Table 4.8. Number of Instances With the Same Influential Word (C-I, I-C)

	C-I	I-C
# of instances with same influential word	14	2
Total	84	59

Table 4.9. Domain of OOV words in the Original Dataset

Type of OOV word	Correct	Incorrect
Relevant to the dataset	224 (32.32%)	88 (12.7%)
Irrelevant to the dataset	254 (36.65%)	127 (18.33%)
Total	478	215

Table 4.10. The most frequent OOV words

Type of OOV word	Correct	Incorrect
Relevant to the dataset	snatching (23), groped (18), inappropriately (17), indecent (14), groping (14)	catcall (14), catcalled (9), misbehaved (8), leh (7), catcalling (6)
Irrelevant to the dataset	rickshaw (14), chowk (13), buttock (8), rajiv (7), quot (5)	safecity (6), rickshaw (4), buea (4), winking (3)

instances. On the other hand, in the I-C case, the model selects the same word as the most influential word in 2 instances out of 59 instances.

4.4 OOV Feature Analysis

The author investigated the characteristics of OOV words in the original dataset concerning instances where the model correctly and incorrectly classified them. This analysis is only subject to the model trained on the original dataset because only the original dataset contains OOV words.

Among 1,375 instances in the test set, 693 instances (50.4%) contain OOV words. Out of 478 instances when the model’s predictions were accurate, 224 occurrences had OOV words that are relevant to the dataset and 254 instances had OOV words that are irrelevant to the dataset. In contrast, out of 215 instances of OOV words, 88 were identified as relevant OOV words and 127 as irrelevant OOV words when the model’s predictions were inaccurate.

The author examined the most frequent OOV words for each scenario: when the model classified an instance correctly and when it misclassified it. The prevalent OOV words differed

between the two cases. In instances where the model accurately classified, the frequently occurring OOV words that are relevant to the dataset included 'snatching,' 'groped,' 'inappropriately,' 'indecent,' and 'groping.' On the other hand, OOV words that are irrelevant to the dataset such as 'rickshaw,' 'chowk,' 'buttok,' 'rajiv,' and 'quot' were common. For cases where the model misclassified, dataset-relevant words like 'catcall,' 'catcalled,' 'misbehaved,' 'leh,' and 'catcalling' were frequently observed, whereas dataset-irrelevant terms such as 'safecity,' 'rickshaw,' 'buea,' and 'winking' appeared frequently. These findings are depicted in Table 4.10, illustrating the prevalent words and their frequencies for each case.

4.5 Summary

In this chapter, the author describes the results of the experiments conducted. In the next chapter, the author will discuss the findings from the results, limitations, and future plans.

5. DISCUSSION, LIMITATIONS AND FUTURE PLAN

In this chapter, the author will provide the interpretation of the results, limitations of our approach, and future plans.

5.1 Discussion

5.1.1 Model Performance

The results indicate that the model trained on the original dataset showed varying performance across different classes with an overall accuracy of 80%. The model demonstrated relatively high recall in the "Commenting" class and the "Groping" class, suggesting that the model accurately identified instances in this class. The recall for the "Ogling" class was 53%, suggesting that the model accurately detected a substantial number of commenting instances but also misclassified instances from other classes as commenting. Furthermore, the "Ogling" class had a relatively low F1-Score (62%), showing that the model had difficulty reliably detecting instances in this class. The "Groping" class showed the highest precision of 82%, demonstrating a strong ability to identify instances belonging to this class correctly.

While the differences were not substantial, the model achieved a slightly lower overall accuracy of 79% when trained on the replaced dataset. Precision and recall values for the "Commenting" and "Groping" classes demonstrated minimal differences of within 2% before and after OOV replacement, indicating relatively consistent results. However, for the "Ogling" class, there was a noticeable decrease (10%) in recall from 53% to 43% after OOV replacement.

Despite the marginal differences observed, the overall trend suggests a slight decline in model performance following OOV replacement. The decrease in recall and F1-Score for the "Ogling" class after OOV replacement implies that the model's ability to correctly identify instances within this class decreased. This could mean that the distribution or properties of the data were changed by replacing OOV terms, which made it harder for the model to correctly identify instances of the "Ogling" class. The minimal changes in precision, recall, and F1-score for the "Commenting" and "Groping" classes suggest that the impact

of OOV replacement on these classes was relatively insignificant. This could indicate that the original dataset provided sufficient information for the model to effectively learn and generalize patterns within these classes, even after the replacement of OOV words. Overall, our results highlight the complex and class-specific impacts of OOV substitution on model performance.

5.1.2 SHAP Analysis

The analysis of influential words through SHAP values provides valuable insights into the model's decision-making process and the impact of OOV replacement on classification results.

The results reveal the distinct patterns of influential words for each class. Words like "commenting," "commented," and "comments," which are associated with harassing behavior and commenting, played a crucial role in correctly classifying instances in the "Commenting" class. On the other hand, instances misclassified within this class were often characterized by the presence of terms like "staring," suggesting challenges in delineating clear boundaries between classes. In the "Ogling" class, similar observations were noted, where words related to visual attention, such as "staring" and "gazing" were influential in correctly classifying instances. However, misclassified instances exhibited more varied items, including "guys" and "touched", indicating potential ambiguity in class distinctions. Similarly, in the "Groping" class, words indicating physical touch, such as "touched" and "touch" were identified as the most frequently occurring influential words in the case of correctly classified instances. When the instance in this class is misclassified, more general words such as "boys", and "pictures" were identified as the most influential words.

After OOV replacement, the analysis showed notable changes in influential words. For instance, in the "Commenting" class, the replacement led to a shift from specific commenting-related terms to more general terms like "and," reflecting a broader scope of influence in the decision-making process. Despite these changes, certain influential words remained consistent across correctly classified instances within each class, highlighting the robustness of these features in capturing class characteristics. However, influential terms that were not

necessarily representative of the classlike "staring" in the "Commenting" class or "pictures" in the "Groping" class were frequently found in misclassified cases.

It indicates the model's ability to capture domain-specific words, but the model exhibited a tendency to misclassify instances when they contained domain-related words that did not clearly belong to one class, or when context was required to discern the content. This highlights the complexity of the classification task in the sexual harassment domain and the need for further refinement in model training and feature selection strategies.

5.1.3 Comparison of Model Prediction on the Original Dataset and Replaced Dataset

Same Results Before and After OOV Replacement (C-C, I-I)

The analysis of instances correctly classified both before and after replacing OOV words reveals the following insights. While a significant number of instances were classified into the correct class both before and after OOV word replacement, there were often differences in the words identified as having the most influence on the model's decisions, as revealed in 4.2. The model exhibited a tendency to select different influential words before and after replacement, regardless of whether it correctly or incorrectly classified the instances, with this tendency being more pronounced in cases of incorrect classification. The presence and quantity of OOV words in the original data were not crucial factors since the number of OOV words in each case widely varied, from instances ranging from those without any OOV words to those containing multiple OOV words.

In case instances were correctly classified into their respective classes (C-C), the author delved deeper into the words the model identified as most influential. After replacing OOV words, the words selected by the model as influential tend to be more domain-related. For example, in Figure 5.1, where the OOV words 'RanchiBhagalpur' and 'Andai' were replaced with more general words like 'Indian' and 'canal', the model selected 'touched' as the most influential word after replacement, whereas 'people' was chosen before replacement. In another example, Figure 5.2, where the OOV word was 'snatching' replaced by 'grab', the word 'grab' became the most influential after replacement, whereas 'chain' was chosen before re-

placement. Furthermore, in the examples provided in Figure 5.3, where there were no OOV words, the model selected 'crowd' and 'man' as the most important words for each sentence before replacement, while 'touched' and 'grabbed' were chosen after replacement. Even in cases where there were no OOV words, the model tended to select domain-related words as the most influential, whether the OOV word was relevant to the domain or not. Additionally, in the "Ogling" class, 'ogling' was mostly replaced by 'gazing'. Before replacement, 'og' or 'ling' were chosen as the most influential words, while 'gazing' was selected after replacement. Although the original model's vocabulary did not include the word 'ogling', it seemed to understand its influence contextually. For sentences containing the word 'staring', the model consistently selected 'staring' as the keyword before and after replacement.

Original: Few bad **people** touched my private pants in RanchiBhagalpur express at Andai railway station.

Replaced: Few bad people **touched** my private parts in Indian express at canal railway station.

Figure 5.1. Instance classified correctly both before and after replacement

Original: **Chain** snatching.

Replaced: Chain **grab**.

Figure 5.2. Instance classified correctly both before and after replacement

I was buying vegetables in the market when a man came in the ^{before} **crowd** ^{after} **touched** me in a wrong way and walked off

I was walking on the street when a ^{before} **man** ^{after} **grabbed** my breast and ran away. He was nowhere to be found.

Figure 5.3. Instance classified correctly both before and after replacement

The experiment showed similar results in cases where the model incorrectly classified the instances both before and after the OOV replacement (I-I). In the example in Figure 5.4, 'idling' was the OOV word replaced by 'wandered'. Before replacement, 'girls' was selected as

the most influential word, whereas after replacement, 'whistling' became the most influential. Even in instances where there were no OOV words, the most influential words chosen by the model before and after replacement differed. In the example in Figure 5.5, the model trained on the original data chose 'while' as the most influential word, while the model trained on the replaced data selected 'whistling'. It was observed that even when the model misclassified the class, the model trained on the replaced data tended to select more domain-related words as the most influential. In addition, when the model incorrectly classified instances both before and after the OOV replacement, in almost all cases it can be seen that both models classified data into the same class.

It can be inferred that in case the model correctly classified the instance, the model demonstrates improved contextual comprehension following the replacement of OOV words, and this pattern persists regardless of the presence or frequency of OOV words. Moreover, it suggests that the model trained on the dataset after OOV replacement has the potential to exhibit enhanced performance and improved ability to generalize. On the other hand, the result suggests that if the model is incorrectly predicted, the classification performance of the model did not change significantly before and after the OOV replacement.

Original: Street guys idling besides the streets, winking and whistling and calling girls.

Replaced: Street guys wandered besides the streets, winking and whistling and calling girls.

Figure 5.4. Instance classified incorrectly both before and after replacement

after before
Whistling while girls go walking from campus to hotel.

Figure 5.5. Instance classified incorrectly both before and after replacement

In Table 4.6 and 4.7, we can see that when the model properly selected keywords related to the class, it classified them correctly, while selecting ambiguous words or words unrelated to the domain resulted in incorrect classification.

Different results before and after replacement (C-I, I-C)

It is shown that cases showing different results between the original dataset and the dataset after OOV word replacement predominantly occur in the 'Groping' class.

In instances where the model initially classified correctly but misclassified after the replacement of OOV words, there were several cases where two models produced different classification outcomes for the same sentences. For instance, as depicted in Figure 5.6, the model trained on the original dataset correctly classified the sentence by selecting "person" as the most influential word, while the model trained on the dataset after OOV replacement misclassified the same sentence by choosing "car" as the most influential word. Similarly, in Figure 5.7, both models selected "me" as the most influential word, with the model trained on the original dataset correctly classifying the sentence while the model trained on the dataset after OOV replacement misclassified it. These instances suggest that OOV replacement may lead to a decrease in model performance in case there's no OOV word in the sentence.

before after
Person stalking in a car.

Figure 5.6. Instance classified correctly before but incorrectly after replacement

before & after
When I was going to tuition class, a guy was stalking me.

Figure 5.7. Instance classified correctly before but incorrectly after replacement

However, when considering instances with spelling errors or multiple OOV words in the original dataset, a different trend was observed. The model after OOV replacement showed better performance in these cases. For example, as shown in Figure 5.8, the model trained on the original dataset misclassified the sentence by selecting "thank" as the most influential word, whereas the model trained on the dataset after OOV replacement correctly classified the sentence by selecting "comments".

Original: I was going home yesterday to my grandma's place when I met this boy that studied in our schoo, I was carrying things in both my hands. I stretthced my hands so he could greet my wrist but instead he hugged me so tight. I then pulled away and he said I looked beautiful and sexy. i was mnot copmfortable with his comments but i said thank you. he later invited me to his house to cook for him but i declined since he was taking the conversation too far and also considaering the way he was looking at me.

Replaced: I was going home yesterday to my grandma's place when I met this boy that studied in our school, I was carrying things in both my hands. I stretched my hands so he could greet my wrist but instead he hugged me so tight. I then pulled away and he said I looked beautiful and sexy. i was not comfortable with his comments but i said thank you. he later invited me to his house to cook for him but i declined since he was taking the conversation too far and also considering the way he was looking at me.

Figure 5.8. Instance classified incorrectly before but correctly after replacement

Furthermore, the author thoroughly examines the cases in which the model consistently chooses the same word as the most influential, both before and after OOV substitution. In Figure 4.8, there are 14 instances for the C-I case and 2 instances for the I-C case. In the C-I case, the majority (12) of them were instances that did not include OOV words. This indicates model after OOV replacement struggles to classify the instance without the OOV word. On the other hand, in the I-C case, the model showed similar behavior before and after OOV replacement. Except for 2 instances, model after replacement generally selects more domain-specific words as the most influential word, leading to accurate predictions.

These findings suggest that, while OOV replacement may have a negative impact on model performance in case there are no OOV words in the sentence, it can improve performance in scenarios including multiple spelling errors or multiple OOV words. From this, it can be inferred that OOV replacement can enhance the contextual understanding of the model and contribute to the model performance when the dataset contains multiple OOV words or words with typos. Therefore, the impact of replacing OOV words on model performance can vary depending on the sentence's structure and context.

5.1.4 OOV Feature Analysis

The distribution of OOV words that are relevant to the dataset and irrelevant to the dataset differs between instances classified correctly and incorrectly by the model. When the model’s predictions were correct, instances frequently included OOV words relevant to the sexual harassment domain, such as ‘snatching,’ ‘groped,’ and ‘inappropriately,’ demonstrating that the model effectively captured domain-specific language patterns. In contrast, cases misclassified by the model had a higher frequency of OOV words that are irrelevant to the dataset, indicating that the model may struggle to understand and contextualize the instances including novel keywords outside the domain.

The analysis of the most frequent OOV words further highlights the distinction between correctly and incorrectly classified instances. In the case where the model correctly classified data, prevalent OOV words that are relevant to the dataset reflect themes relevant to the classification task, such as instances of harassment or inappropriate behavior. Prevalent OOV words that are irrelevant to the dataset were the words that indicate taxi, the part of the body, and the name of the place. On the other hand, misclassified cases frequently include OOV words that are relevant to the dataset and were ambiguous to classify into one class, such as ‘catcall’ and ‘misbehaved’, and this leads the model to become confused or misread novel phrases. OOV words that are irrelevant to the dataset contain words like ‘Safecity’ and ‘winking’.

5.2 Limitations

While this study provides insights into the impact of OOV words in the sexual harassment domain, several limitations should be considered. Firstly, the findings are specific to the SafeCity dataset, which can limit their generalizability to different datasets in sexual harassment domains. Furthermore, factors like the OOV replacement method, the quality of replaced words, and their semantic alignment with the OOV terms can affect how effective OOV replacement is. Moreover, this study only focuses on a single LLM model and its classification ability, so the findings may vary depending on the type of LLM and task.

Further research is necessary to explore the biases in the model training procedure and the interpretation of misclassification.

5.3 Future Plan

In the future, the author aims to address several key limitations and expand the range of the study to get more generalizable results. First, the author recognizes the importance of dataset diversity and intends to broaden our dataset collection to compare the results within different datasets. By comparing various kinds of sexual harassment story datasets, the author will be able to get an enhanced and generalized understanding of the impact of OOV words in classification tasks within this domain. Additionally, the author plans to expand the type of LLM and compare each model to get a broader perspective. Conducting comparative studies will provide insights into the strengths and weaknesses of different LLMs, as well as their input sensitivity to the OOV words in narrow domains.

5.4 Summary

This section provides the discussion, limitations, and future plan of this study.

REFERENCES

- [1] M. Chawki and Y. El Shazly, “Online sexual harassment: Issues & solutions,” *J. Intell. Prop. Info. Tech. & Elec. Com. L.*, vol. 4, p. 71, 2013.
- [2] S. Welsh, J. Carr, B. MacQuarrie, and A. Huntley, “im not thinking of it as sexual harassment understanding harassment across race and citizenship,” *Gender & Society*, vol. 20, no. 1, pp. 87–107, 2006.
- [3] H. Fujita *et al.*, “Prompt sensitivity of language model for solving programming problems,” in *New Trends in Intelligent Software Methodologies, Tools and Techniques: Proceedings of the 21st International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT_22)*, IOS Press, vol. 355, 2022, p. 346.
- [4] J. Wei, X. Wang, D. Schuurmans, *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.
- [5] Pew Research Center, “The state of online harassment,” en-US, Washington, D.C., Tech. Rep., Jan. 2021. [Online]. Available: <https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/>.
- [6] N. Henry and A. Powell, “Embodied harms: Gender, shame, and technology-facilitated sexual violence,” *Violence against women*, vol. 21, no. 6, pp. 758–779, 2015.
- [7] M. Gámez-Guadix, C. Almendros, E. Borrajo, and E. Calvete, “Prevalence and association of sexting and online sexual victimization among spanish adults,” *Sexuality Research and Social Policy*, vol. 12, pp. 145–154, 2015.
- [8] J. K. Biber, D. Doverspike, D. Baznik, A. Cober, and B. A. Ritter, “Sexual harassment in online communications: Effects of gender and discourse medium,” *CyberPsychology & Behavior*, vol. 5, no. 1, pp. 33–42, 2002.
- [9] N. Henry and A. Powell, “Technology-facilitated sexual violence: A literature review of empirical research,” *Trauma, violence, & abuse*, vol. 19, no. 2, pp. 195–208, 2018.
- [10] L. Mahr, J. Montgomery, and M. Ramírez, “Interpretation Issues In Sexual Harassment Cases On Behalf Of Female Farmworkers,” en, 2007.

- [11] J. Wei, Y. Tay, R. Bommasani, *et al.*, “Emergent abilities of large language models,” *arXiv preprint arXiv:2206.07682*, 2022.
- [12] Y. Chang, X. Wang, J. Wang, *et al.*, “A survey on evaluation of large language models,” *arXiv preprint arXiv:2307.03109*, 2023.
- [13] D. Hovy and S. L. Spruit, “The social impact of natural language processing,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016, pp. 591–598.
- [14] A. G. Chowdhury, R. Sawhney, R. Shah, and D. Mahata, “# Youtoo? detection of personal recollections of sexual harassment on social media,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2527–2537.
- [15] M. Yan and X. Luo, “Bert-based detection of sexual harassment in dialogues,” in *Proceedings of the 2021 5th International Conference on Computer Science and Artificial Intelligence*, 2021, pp. 359–364.
- [16] H. Inan, K. Upasani, J. Chi, *et al.*, “Llama guard: Llm-based input-output safeguard for human-ai conversations,” *arXiv preprint arXiv:2312.06674*, 2023.
- [17] P. Basu, T. Singha Roy, S. Tiwari, and S. Mehta, “Cyberpolice: Classification of cyber sexual harassment,” in *Progress in Artificial Intelligence: 20th EPIA Conference on Artificial Intelligence, EPIA 2021, Virtual Event, September 7–9, 2021, Proceedings 20*, Springer, 2021, pp. 701–714.
- [18] M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr, “Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting,” *arXiv preprint arXiv:2310.11324*, 2023.
- [19] K. Misra, A. Ettinger, and J. T. Rayz, “Exploring bert’s sensitivity to lexical cues using tests from semantic priming,” *arXiv preprint arXiv:2010.03010*, 2020.
- [20] S. Karlekar and M. Bansal, “Safecity: Understanding diverse forms of sexual harassment personal stories,” *arXiv preprint arXiv:1809.04739*, 2018.
- [21] M. U. Hadi, R. Qureshi, A. Shah, *et al.*, “A survey on large language models: Applications, challenges, limitations, and practical usage,” *TechRxiv*, 2023.

- [22] E. D. Liddy, “Natural language processing,” 2001.
- [23] P. Kolachina, M. Riedl, and C. Biemann, “Replacing oov words for dependency parsing with distributional semantics,” in *Proceedings of the 21st Nordic conference on computational linguistics*, 2017, pp. 11–19.
- [24] J. V. Lochter, R. M. Silva, and T. A. Almeida, “Multi-level out-of-vocabulary words handling approach,” *Knowledge-Based Systems*, vol. 251, p. 108911, 2022.
- [25] H. Naveed, A. U. Khan, S. Qiu, *et al.*, “A comprehensive overview of large language models,” *arXiv preprint arXiv:2307.06435*, 2023.
- [26] H. Huang, O. Zheng, D. Wang, *et al.*, “Chatgpt for shaping the future of dentistry: The potential of multi-modal large language model,” *International Journal of Oral Science*, vol. 15, no. 1, p. 29, 2023.
- [27] N. Du, Y. Huang, A. M. Dai, *et al.*, “Glam: Efficient scaling of language models with mixture-of-experts,” in *International Conference on Machine Learning*, PMLR, 2022, pp. 5547–5569.
- [28] W. ODonohue, K. Downs, and E. A. Yeater, “Sexual harassment: A review of the literature,” *Aggression and Violent Behavior*, vol. 3, no. 2, pp. 111–128, 1998.
- [29] *Sex discrimination*, <https://www.eeoc.gov/laws/guidance/sex-discrimination>.
- [30] T. P. Sbraga and W. O’donohue, “Sexual harassment,” *Annual review of sex research*, vol. 11, no. 1, pp. 258–285, 2000.
- [31] J. L. Berdahl, “Harassment based on sex: Protecting social status in the context of gender hierarchy,” *Academy of management review*, vol. 32, no. 2, pp. 641–658, 2007.
- [32] C. A. MacKinnon and C. A. MacKinnon, *Sexual harassment of working women: A case of sex discrimination*. Yale University Press, 1979, vol. 19.
- [33] N. D. Betts and G. C. Newman, “Defining the issue: Sexual harassment in college and university life,” *Contemporary Education*, vol. 54, no. 1, p. 48, 1982.
- [34] L. F. Fitzgerald, S. Swan, and V. J. Magley, “But was it really sexual harassment?: Legal, behavioral, and psychological definitions of the workplace victimization of women.,” 1997.

- [35] P. McDonald, “Workplace sexual harassment 30 years on: A review of the literature,” *International Journal of Management Reviews*, vol. 14, no. 1, pp. 1–17, 2012.
- [36] Pew Research Center, “Online harassment,” en-US, Washington, D.C., Tech. Rep., Oct. 2014. [Online]. Available: <https://www.pewresearch.org/internet/2014/10/22/online-harassment/>.
- [37] A. Barak, “Sexual harassment on the internet,” *Social science computer review*, vol. 23, no. 1, pp. 77–92, 2005.
- [38] M. L. Pittaro, “Cyber stalking: An analysis of online harassment and intimidation,” *International journal of cyber criminology*, vol. 1, no. 2, pp. 180–197, 2007.
- [39] H. Whittle, C. Hamilton-Giachritsis, A. Beech, and G. Collings, “A review of online grooming: Characteristics and concerns,” *Aggression and violent behavior*, vol. 18, no. 1, pp. 62–70, 2013.
- [40] J. A. Pater, M. K. Kim, E. D. Mynatt, and C. Fiesler, “Characterizations of online harassment: Comparing policies across social media platforms,” in *Proceedings of the 2016 ACM International Conference on Supporting Group Work*, 2016, pp. 369–374.
- [41] M. Duggan, “Online harassment 2017,” 2017.
- [42] A. Lenhart, M. Ybarra, K. Zickuhr, and M. Price-Feeney, *Online harassment, digital abuse, and cyberstalking in America*. Data and Society Research Institute, 2016.
- [43] A. Powell and N. Henry, *Sexual violence in a digital age*. Springer, 2017.
- [44] D. Finkelhor, K. J. Mitchell, and J. Wolak, “Online victimization: A report on the nation’s youth,” 2000.
- [45] A. Slaughter and E. Newman, “New frontiers: Moving beyond cyberbullying to define online harassment,” *Journal of Online Trust and Safety*, vol. 1, no. 2, 2022.
- [46] D. P. Ford, “Virtual harassment: Media characteristics’ role in psychological health,” *Journal of Managerial Psychology*, vol. 28, no. 4, pp. 408–428, 2013.

- [47] T. T. Ojanen, P. Boonmongkon, R. Samakkeekarom, *et al.*, “Investigating online harassment and offline violence among young people in thailand: Methodological approaches, lessons learned,” *Culture, health & sexuality*, vol. 16, no. 9, pp. 1097–1112, 2014.
- [48] T. Van Laer, “The means to justify the end: Combating cyber harassment in social media,” *Journal of Business Ethics*, vol. 123, no. 1, pp. 85–98, 2014.
- [49] D. E. Penza, “The unstoppable intrusion: The unique effect of online harassment and what the united states can ascertain from other countries’ attempts to prevent it,” *Cornell Int’l LJ*, vol. 51, p. 297, 2018.
- [50] S. Einarsen, H. Hoel, and G. Notelaers, “Measuring exposure to bullying and harassment at work: Validity, factor structure and psychometric properties of the negative acts questionnaire-revised,” *Work & stress*, vol. 23, no. 1, pp. 24–44, 2009.
- [51] K. M. Rospenda, J. A. Richman, and C. A. Shannon, “Prevalence and mental health correlates of harassment and discrimination in the workplace: Results from a national study,” *Journal of interpersonal violence*, vol. 24, no. 5, pp. 819–843, 2009.
- [52] F. Angela, R.-d. María-Luisa, N. Annalaura, and M. Ersilia, “Online sexual harassment in adolescence: A scoping review,” *Sexuality Research and Social Policy*, pp. 1–20, 2023.
- [53] J. A. Hamilton, S. W. Alagna, L. S. King, and C. Lloyd, “The emotional consequences of gender-based abuse in the workplace: New counseling programs for sex discrimination,” *Women & Therapy*, vol. 6, no. 1-2, pp. 155–182, 1987.
- [54] J. Salisbury, A. B. Ginorio, H. Remick, and D. M. Stringer, “Counseling victims of sexual harassment.,” *Psychotherapy: Theory, Research, Practice, Training*, vol. 23, no. 2, p. 316, 1986.
- [55] P. Crull, “Stress effects of sexual harassment on the job: Implications for counseling,” *American Journal of Orthopsychiatry*, vol. 52, no. 3, pp. 539–544, 1982.
- [56] P. H. Loy and L. P. Stewart, “The extent and effects of the sexual harassment of working women,” *Sociological focus*, vol. 17, no. 1, pp. 31–43, 1984.
- [57] M. P. Rowe, “Dealing with sexual harassment.,” *Harvard Business Review*, vol. 59, no. 3, pp. 42–46, 1981.

- [58] M. P. Koss, “Changed lives: The psychological impact of sexual harassment,” *Ivory power: Sexual harassment on campus*, pp. 73–92, 1990.
- [59] E. Reed, A. Wong, and A. Raj, “Cyber sexual harassment: A summary of current measures and implications for future research,” *Violence against women*, vol. 26, no. 12-13, pp. 1727–1740, 2020.
- [60] M. L. Ybarra, “Linkages between depressive symptomatology and internet harassment among young regular internet users,” *CyberPsychology & Behavior*, vol. 7, no. 2, pp. 247–257, 2004.
- [61] J. A. Scarduzio, S. E. Sheff, and M. Smith, “Coping and sexual harassment: How victims cope across multiple settings,” *Archives of Sexual Behavior*, vol. 47, pp. 327–340, 2018.
- [62] C. Hill and H. Kearl, *Crossing the Line: Sexual Harassment at School*. ERIC, 2011.
- [63] L. A. Melander, “College students’ perceptions of intimate partner cyber harassment,” *Cyberpsychology, behavior, and social networking*, vol. 13, no. 3, pp. 263–268, 2010.
- [64] S. Ståhl and I. Denhag, “Online and offline sexual harassment associations of anxiety and depression in an adolescent sample,” *Nordic journal of psychiatry*, vol. 75, no. 5, pp. 330–335, 2021.
- [65] A. G. Chowdhury, R. Sawhney, P. Mathur, D. Mahata, and R. R. Shah, “Speak up, fight back! detection of social media disclosures of sexual harassment,” in *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Student research workshop*, 2019, pp. 136–146.
- [66] G. B. Dahl and M. M. Knepper, “Why is workplace sexual harassment underreported? the value of outside options amid the threat of retaliation,” National Bureau of Economic Research, Tech. Rep., 2021.
- [67] S. J. Aguilar and C. Baek, “Sexual harassment in academe is underreported, especially by students in the life and physical sciences,” *PloS one*, vol. 15, no. 3, e0230312, 2020.
- [68] J. Hersch, “Sexual harassment in the workplace,” *IZA World of Labor*, 2015.

- [69] N. A. Hamzah and B. N. Dhannoon, “The detection of sexual harassment and chat predators using artificial neural network,” *Karbala International Journal of Modern Science*, vol. 7, no. 4, p. 6, 2021.
- [70] E. Alawneh, M. Al-Fawa’reh, M. T. Jafar, and M. Al Fayoumi, “Sentiment analysis-based sexual harassment detection using machine learning techniques,” in *2021 international symposium on electronics and smart devices (ISESD)*, IEEE, 2021, pp. 1–6.
- [71] M. Saeidi, S. B. da S. Sousa, E. Milios, N. Zeh, and L. Berton, “Categorizing online harassment on twitter,” in *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, Springer, 2020, pp. 283–297.
- [72] Y. Liu, Q. Li, X. Liu, Q. Zhang, and L. Si, “Sexual harassment story classification and key information identification,” in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 2385–2388.
- [73] T. L. Nikmah, M. Z. Ammar, Y. R. Allatif, R. M. P. Husna, P. A. Kurniasari, and A. S. Bahri, “Comparison of lstm, svm, and naive bayes for classifying sexual harassment tweets,” *Journal of Soft Computing Exploration*, vol. 3, no. 2, pp. 131–137, 2022.
- [74] A. S. A. Al-Katheri and M. M. Siraj, “Classification of sexual harassment on facebook using term weighting schemes,” *International Journal of Innovative Computing*, vol. 8, no. 1, 2018.
- [75] M. Rezvan, S. Shekarpour, L. Balasuriya, K. Thirunarayan, V. L. Shalin, and A. Sheth, “A quality type-aware annotated corpus and lexicon for harassment research,” in *Proceedings of the 10th acm conference on web science*, 2018, pp. 33–36.
- [76] U. Bretschneider, T. Wöhner, and R. Peters, “Detecting online harassment in social networks,” 2014.
- [77] I. Espinoza and F. Weiss, “Detection of harassment on twitter with deep learning techniques,” in *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, Springer, 2020, pp. 307–313.
- [78] M. Arslan, M. Guzel, M. Demirci, and S. Ozdemir, “Smote and gaussian noise based sensor data augmentation,” in *2019 4th International Conference on Computer Science and Engineering (UBMK)*, IEEE, 2019, pp. 1–5.

- [79] C. Karatsalos and Y. Panagiotakis, “Attention-based method for categorizing different types of online harassment language,” in *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, Springer, 2020, pp. 321–330.
- [80] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [81] J. Gao and C.-Y. Lin, *Introduction to the special issue on statistical language modeling*, 2004.
- [82] W. X. Zhao, K. Zhou, J. Li, *et al.*, “A survey of large language models,” *arXiv preprint arXiv:2303.18223*, 2023.
- [83] A. Nadas, “Estimation of probabilities in the language model of the ibm speech recognition system,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 4, pp. 859–861, 1984.
- [84] S. F. Chen and J. Goodman, “An empirical study of smoothing techniques for language modeling,” *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999.
- [85] T. Brown, B. Mann, N. Ryder, *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [86] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [87] A. Chowdhery, S. Narang, J. Devlin, *et al.*, “Palm: Scaling language modeling with pathways,” *arXiv preprint arXiv:2204.02311*, 2022.
- [88] R. Taylor, M. Kardas, G. Cucurull, *et al.*, “Galactica: A large language model for science,” *arXiv preprint arXiv:2211.09085*, 2022.
- [89] H. Touvron, T. Lavril, G. Izacard, *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [90] M. Shanahan, “Talking about large language models,” *arXiv preprint arXiv:2212.03551*, 2022.

- [91] Y. Bengio, R. Ducharme, and P. Vincent, “A neural probabilistic language model,” *Advances in neural information processing systems*, vol. 13, 2000.
- [92] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Advances in neural information processing systems*, vol. 26, 2013.
- [93] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [94] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” *Advances in neural information processing systems*, vol. 27, 2014.
- [95] M.-T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” *arXiv preprint arXiv:1508.04025*, 2015.
- [96] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.
- [97] Y. Xia, F. Tian, L. Wu, *et al.*, “Deliberation networks: Sequence generation beyond one-pass decoding,” *Advances in neural information processing systems*, vol. 30, 2017.
- [98] K. Xu, J. Ba, R. Kiros, *et al.*, “Show, attend and tell: Neural image caption generation with visual attention,” in *International conference on machine learning*, PMLR, 2015, pp. 2048–2057.
- [99] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2016, pp. 4945–4949.
- [100] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [101] M. Sundermeyer, R. Schlüter, and H. Ney, “Lstm neural networks for language modeling,” in *Thirteenth annual conference of the international speech communication association*, 2012.

- [102] G. Kurata, B. Ramabhadran, G. Saon, and A. Sethy, “Language modeling with highway lstm,” in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2017, pp. 244–251.
- [103] T. Mikolov and G. Zweig, “Context dependent recurrent neural network language model,” in *2012 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2012, pp. 234–239.
- [104] J. Xiao and Z. Zhou, “Research progress of rnn language model,” in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, IEEE, 2020, pp. 1285–1288.
- [105] S. Jean, K. Cho, R. Memisevic, and Y. Bengio, “On using very large target vocabulary for neural machine translation,” *arXiv preprint arXiv:1412.2007*, 2014.
- [106] X. Li, T. Qin, J. Yang, and T.-Y. Liu, “Lightrnn: Memory and computation-efficient recurrent neural networks,” *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [107] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in *International conference on machine learning*, Pmlr, 2013, pp. 1310–1318.
- [108] N. Park and S. Kim, “How do vision transformers work?” *arXiv preprint arXiv:2202.06709*, 2022.
- [109] T. Wolf, L. Debut, V. Sanh, *et al.*, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.
- [110] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [111] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, “Transformer-xl: Attentive language models beyond a fixed-length context,” *arXiv preprint arXiv:1901.02860*, 2019.
- [112] S. Zhang, S. Roller, N. Goyal, *et al.*, “Opt: Open pre-trained transformer language models,” *arXiv preprint arXiv:2205.01068*, 2022.

- [113] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, “Improving language understanding by generative pre-training,” 2018.
- [114] Y. Liu, M. Ott, N. Goyal, *et al.*, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [115] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 2019.
- [116] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” *arXiv preprint arXiv:1909.11942*, 2019.
- [117] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” *Advances in neural information processing systems*, vol. 32, 2019.
- [118] Y. Sun, S. Wang, Y. Li, *et al.*, “Ernie 2.0: A continual pre-training framework for language understanding,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 8968–8975.
- [119] C. Raffel, N. Shazeer, A. Roberts, *et al.*, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.
- [120] M. Lewis, Y. Liu, N. Goyal, *et al.*, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” *arXiv preprint arXiv:1910.13461*, 2019.
- [121] J. Kaplan, S. McCandlish, T. Henighan, *et al.*, “Scaling laws for neural language models,” *arXiv preprint arXiv:2001.08361*, 2020.
- [122] W. Zeng, X. Ren, T. Su, *et al.*, “Pangu-alpha: Large-scale autoregressive pretrained chinese language models with auto-parallel computation,” *arXiv preprint arXiv:2104.12369*, 2021.
- [123] D. Lepikhin, H. Lee, Y. Xu, *et al.*, “Gshard: Scaling giant models with conditional computation and automatic sharding,” *arXiv preprint arXiv:2006.16668*, 2020.

- [124] W. Fedus, B. Zoph, and N. Shazeer, “Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity,” *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 5232–5270, 2022.
- [125] A. Chowdhery, S. Narang, J. Devlin, *et al.*, “Palm: Scaling language modeling with pathways,” *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.
- [126] L. Martin, B. Muller, P. J. O. Suárez, *et al.*, “Camembert: A tasty french language model,” *arXiv preprint arXiv:1911.03894*, 2019.
- [127] S. Lee, H. Jang, Y. Baik, S. Park, and H. Shin, “Kr-bert: A small-scale korean-specific language model,” *arXiv preprint arXiv:2008.03979*, 2020.
- [128] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” *arXiv preprint arXiv:1801.06146*, 2018.
- [129] Y. Meng, Y. Zhang, J. Huang, *et al.*, “Text classification using label names only: A language model self-training approach,” *arXiv preprint arXiv:2010.07245*, 2020.
- [130] N. Kant, R. Puri, N. Yakovenko, and B. Catanzaro, “Practical text classification with large pre-trained language models,” *arXiv preprint arXiv:1812.01207*, 2018.
- [131] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, “A survey of controllable text generation using transformer-based pre-trained language models,” *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–37, 2023.
- [132] J. Li, T. Tang, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, “Pretrained language models for text generation: A survey,” *arXiv preprint arXiv:2201.05273*, 2022.
- [133] L. F. Ribeiro, M. Schmitt, H. Schütze, and I. Gurevych, “Investigating pretrained language models for graph-to-text generation,” *arXiv preprint arXiv:2007.08426*, 2020.
- [134] Y. Zhu, H. Yuan, S. Wang, *et al.*, “Large language models for information retrieval: A survey,” *arXiv preprint arXiv:2308.07107*, 2023.
- [135] V. Jeronimo, L. Bonifacio, H. Abonizio, *et al.*, “Inpars-v2: Large language models as efficient dataset generators for information retrieval,” *arXiv preprint arXiv:2301.01820*, 2023.

- [136] Y. Dan, Z. Lei, Y. Gu, *et al.*, “Educhat: A large-scale language model-based chatbot system for intelligent education,” *arXiv preprint arXiv:2308.02773*, 2023.
- [137] G. Campagna, S. Xu, M. Moradshahi, R. Socher, and M. S. Lam, “Genie: A generator of natural language semantic parsers for virtual assistant commands,” in *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2019, pp. 394–410.
- [138] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean, “Large language models in machine translation,” 2007.
- [139] C. Lyu, J. Xu, and L. Wang, “New trends in machine translation using large language models: Case examples with chatgpt,” *arXiv preprint arXiv:2305.01181*, 2023.
- [140] Y. Moslem, R. Haque, and A. Way, “Adaptive machine translation with large language models,” *arXiv preprint arXiv:2301.13294*, 2023.
- [141] T. Gupta, M. Zaki, N. A. Krishnan, and Mausam, “Matscibert: A materials domain language model for text mining and information extraction,” *npj Computational Materials*, vol. 8, no. 1, p. 102, 2022.
- [142] T. Zhang, Z. Cai, C. Wang, M. Qiu, B. Yang, and X. He, “Smedbert: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining,” *arXiv preprint arXiv:2108.08983*, 2021.
- [143] J. Lee, W. Yoon, S. Kim, *et al.*, “Biobert: A pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [144] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, “Finbert: A pre-trained financial language representation model for financial text mining,” in *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, 2021, pp. 4513–4519.
- [145] L. Tang, Z. Sun, B. Idnay, *et al.*, “Evaluating large language models on medical evidence summarization,” *npj Digital Medicine*, vol. 6, no. 1, p. 158, 2023.
- [146] J. Pilault, R. Li, S. Subramanian, and C. Pal, “On extractive and abstractive neural document summarization with transformer language models,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 9308–9319.

- [147] Y. Du, Q. Li, L. Wang, and Y. He, “Biomedical-domain pre-trained language model for extractive summarization,” *Knowledge-Based Systems*, vol. 199, p. 105 964, 2020.
- [148] P. Yang, W. Li, and G. Zhao, “Language model-driven topic clustering and summarization for news articles,” *IEEE Access*, vol. 7, pp. 185 506–185 519, 2019.
- [149] S. Toshniwal, A. Kannan, C.-C. Chiu, Y. Wu, T. N. Sainath, and K. Livescu, “A comparison of techniques for language model integration in encoder-decoder speech recognition,” in *2018 IEEE spoken language technology workshop (SLT)*, IEEE, 2018, pp. 369–375.
- [150] D. Vergyri, K. Kirchhoff, K. Duh, and A. Stolcke, “Morphology-based language modeling for arabic speech recognition,” in *INTERSPEECH*, vol. 4, 2004, pp. 2245–2248.
- [151] Y. Fathullah, C. Wu, E. Lakomkin, *et al.*, “Prompting large language models with speech recognition abilities,” *arXiv preprint arXiv:2307.11795*, 2023.
- [152] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, “Large language models in medicine,” *Nature medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.
- [153] S. Dathathri, A. Madotto, J. Lan, *et al.*, “Plug and play language models: A simple approach to controlled text generation,” *arXiv preprint arXiv:1912.02164*, 2019.
- [154] L. Sha, “Gradient-guided unsupervised lexically constrained text generation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 8692–8703.
- [155] Z. Song, X. Zheng, L. Liu, M. Xu, and X.-J. Huang, “Generating responses with a specific emotion in dialog,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3685–3695.
- [156] Z. Lin and M. O. Riedl, “Plug-and-blend: A framework for plug-and-play controllable story generation with sketches,” in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 17, 2021, pp. 58–65.
- [157] Y. Wang, I. Wood, S. Wan, M. Dras, and M. Johnson, “Mention flags (mf): Constraining transformer-based text generators,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 103–113.

- [158] L. Bonifacio, H. Abonizio, M. Fadaee, and R. Nogueira, “Inpars: Data augmentation for information retrieval using large language models,” *arXiv preprint arXiv:2202.05144*, 2022.
- [159] H. Decker, K. Trang, J. Ramirez, *et al.*, “Large language model- based chatbot vs surgeon-generated informed consent documentation for common procedures,” *JAMA Network Open*, vol. 6, no. 10, e2336997–e2336997, 2023.
- [160] I. A. Bernstein, Y. V. Zhang, D. Govil, *et al.*, “Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions,” *JAMA Network Open*, vol. 6, no. 8, e2330320–e2330320, 2023.
- [161] A. A. Birkun and A. Gautam, “Large language model-based chatbot as a source of advice on first aid in heart attack,” *Current Problems in Cardiology*, p. 102 048, 2023.
- [162] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, “Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model,” *arXiv preprint arXiv:1906.01749*, 2019.
- [163] A. Bajaj, P. Dangati, K. Krishna, *et al.*, “Long document summarization in a low resource setting using pretrained language models,” *arXiv preprint arXiv:2103.00751*, 2021.
- [164] Q. Grail, J. Perez, and E. Gaussier, “Globalizing bert-based transformer architectures for long document summarization,” in *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main volume*, 2021, pp. 1792–1810.
- [165] M. Shao, B. Chen, S. Jancheska, *et al.*, “An empirical evaluation of llms for solving offensive security challenges,” *arXiv preprint arXiv:2402.11814*, 2024.
- [166] K. Guo, A. Hu, J. Mu, *et al.*, “An investigation of large language models for real-world hate speech detection,” *arXiv preprint arXiv:2401.03346*, 2024.
- [167] K.-L. Chiu, A. Collins, and R. Alexander, “Detecting hate speech with gpt-3,” *arXiv preprint arXiv:2103.12407*, 2021.
- [168] S. Paul and S. Saha, “Cyberbert: Bert for cyberbullying identification: Bert for cyberbullying identification,” *Multimedia Systems*, vol. 28, no. 6, pp. 1897–1904, 2022.

- [169] M. Zhang, J. He, T. Ji, and C.-T. Lu, “Don’t go to extremes: Revealing the excessive sensitivity and calibration limitations of llms in implicit hate speech detection,” *arXiv preprint arXiv:2402.11406*, 2024.
- [170] O. Adams, A. Makarucha, G. Neubig, S. Bird, and T. Cohn, “Cross-lingual word embeddings for low-resource language modeling,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 937–947.
- [171] D. Wang, “Out-of-vocabulary spoken term detection,” 2010.
- [172] L. Chen, G. Varoquaux, and F. M. Suchanek, “Imputing out-of-vocabulary embeddings with love makes language models robust with little cost,” *arXiv preprint arXiv:2203.07860*, 2022.
- [173] J. Botha and P. Blunsom, “Compositional morphology for word representations and language modelling,” in *International Conference on Machine Learning*, PMLR, 2014, pp. 1899–1907.
- [174] P. Bhatia, R. Guthrie, and J. Eisenstein, “Morphological priors for probabilistic neural word embeddings,” *arXiv preprint arXiv:1608.01056*, 2016.
- [175] Y. Pinter, R. Guthrie, and J. Eisenstein, “Mimicking word embeddings using subword rnns,” *arXiv preprint arXiv:1707.06961*, 2017.
- [176] W. Ling, T. Luís, L. Marujo, *et al.*, “Finding function in form: Compositional character models for open vocabulary word representation,” *arXiv preprint arXiv:1508.02096*, 2015.
- [177] Y. Kim, Y. Jernite, D. Sontag, and A. Rush, “Character-aware neural language models,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016.
- [178] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, “Charagram: Embedding words and sentences via character n-grams,” *arXiv preprint arXiv:1607.02789*, 2016.
- [179] E. Vylomova, T. Cohn, X. He, and G. Haffari, “Word representation models for morphologically rich languages in neural machine translation,” *arXiv preprint arXiv:1606.04217*, 2016.

- [180] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.
- [181] Z. Hu, T. Chen, K.-W. Chang, and Y. Sun, “Few-shot representation learning for out-of-vocabulary words,” *arXiv preprint arXiv:1907.00505*, 2019.
- [182] N. Garneau, J.-S. Leboeuf, and L. Lamontagne, “Predicting and interpreting embeddings for out of vocabulary words in downstream tasks,” *arXiv preprint arXiv:1903.00724*, 2019.
- [183] R. F. Astudillo, S. Amir, W. Ling, M. J. Silva, and I. Trancoso, “Learning word representations from scarce and noisy data with embedding subspaces,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1074–1084.
- [184] M. Pota, F. Marulli, M. Esposito, G. De Pietro, and H. Fujita, “Multilingual pos tagging by a composite deep architecture based on character-level features and on-the-fly enriched word embeddings,” *Knowledge-Based Systems*, vol. 164, pp. 309–323, 2019.
- [185] O. Kwon, D. Kim, S.-R. Lee, J. Choi, and S. Lee, “Handling out-of-vocabulary problem in hangeul word embeddings,” in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 3213–3221.
- [186] B. Plank, A. Søgaard, and Y. Goldberg, “Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss,” *arXiv preprint arXiv:1604.05529*, 2016.
- [187] Y. Belinkov and Y. Bisk, “Synthetic and natural noise both break neural machine translation,” *arXiv preprint arXiv:1711.02173*, 2017.
- [188] M. Khodak, N. Saunshi, Y. Liang, T. Ma, B. Stewart, and S. Arora, “A la carte embedding: Cheap but effective induction of semantic feature vectors,” *arXiv preprint arXiv:1805.05388*, 2018.
- [189] B. Edizel, A. Piktus, P. Bojanowski, R. Ferreira, E. Grave, and F. Silvestri, “Misspelling oblivious word embeddings,” *arXiv preprint arXiv:1905.09755*, 2019.

- [190] V. Malykh, T. Khakhulin, and V. Logacheva, “Robust word vectors: Context-informed embeddings for noisy texts,” *Journal of Mathematical Sciences*, pp. 1–14, 2023.
- [191] H. E. Boukkouri, O. Ferret, T. Lavergne, H. Noji, P. Zweigenbaum, and J. Tsujii, “Characterbert: Reconciling elmo and bert for word-level open-vocabulary representations from characters,” *arXiv preprint arXiv:2010.10392*, 2020.
- [192] Y. Chae and T. Davidson, “Large language models for text classification: From zero-shot learning to fine-tuning,” *Open Science Foundation*, 2023.
- [193] *Datamuse*, Available: <https://www.datamuse.com/api/>, [Online], Accessed: February 2024.
- [194] *Symspellpy*, Available: <https://symspellpy.readthedocs.io/>, [Online], Accessed: February 2024.
- [195] C. Wang, P. Nulty, and D. Lillis, “A comparative study on word embeddings in deep learning for text classification,” in *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, 2020, pp. 37–46.
- [196] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.