A close-up of a logo

Description automatically generated

**MASTER OF ARTIFICIAL INTELLIGENCE IN BUSINESS PROGRAM**

**Assignment-2**

**MAIB -Natural Language and Conversational Systems with Business Applications**

**Jawwad Ahmed**

Contents

[1. Introduction 3](#_Toc173348036)

[2. Literature Review 4](#_Toc173348037)

[2.1. Traditional Machine Learning for Hate Speech Detection 4](#_Toc173348038)

[2.2. Recurrent Neural Networks (RNNs) and LSTMs 4](#_Toc173348039)

[2.3 Large Language Models (LLMs) for Hate Speech Detection 5](#_Toc173348040)

[2.4 Summary of Cited Papers 5](#_Toc173348041)

[3. Data Collection and Pre-processing 6](#_Toc173348042)

[3.1 Toxic Comment Classification Dataset 6](#_Toc173348043)

[3.2 HateXplain Dataset 7](#_Toc173348044)

[3.3 Data Split 8](#_Toc173348045)

[4.1 Traditional Machine Learning Models 8](#_Toc173348046)

[4.2 Deep Learning Model: Long Short-Term Memory (LSTM) 9](#_Toc173348047)

[4.3 Large Language Models (LLMs) 11](#_Toc173348048)

[5. Results and Discussion 12](#_Toc173348049)

[5.1. Model Performance Overview 12](#_Toc173348050)

[5.2 Analysis and Business Applications 13](#_Toc173348051)

[6. Conclusion 15](#_Toc173348052)

[7. References 17](#_Toc173348053)

## 1. Introduction

The proliferation of online platforms and social media has revolutionized communication, fostering connections and enabling global information exchange. However, this digital landscape has also witnessed a disturbing rise in hate speech – abusive or discriminatory language targeting individuals or groups based on characteristics like race, ethnicity, religion, gender, sexual orientation, or other protected attributes. The spread of hate speech online not only inflicts emotional harm on its victims but also contributes to a toxic online environment, discourages participation, and can even incite real-world violence.

Combating hate speech online is a complex and multifaceted challenge. Traditional moderation methods often rely heavily on human review, which is time-consuming, expensive, and prone to inconsistencies. As a result, there is a growing need for accurate and efficient automated approaches to detect and mitigate hate speech. Natural Language Processing (NLP), a branch of artificial intelligence focused on enabling computers to understand and process human language, offers promising avenues for addressing this problem.

This project focuses on comparing the performance of various NLP techniques for hate speech detection, aiming to identify the most suitable approaches for different business contexts. We evaluate a range of models, including traditional machine learning algorithms (Logistic Regression, Random Forest), a deep learning architecture (Long Short-Term Memory networks - LSTM), and large language models (BERT, GPT-3, GPT-2). Our analysis explores their performance on two distinct datasets: the Toxic Comment Classification dataset and the HateXplain dataset, examining their accuracy, precision, and recall in classifying hate speech.

This report is structured as follows: Section 2 provides a concise overview of existing research and methods in hate speech detection. Section 3 details the datasets employed in our study, outlining the data collection and pre-processing steps. Section 4 explains the selected models, their architectures, and the training process. Section 5 presents a comprehensive analysis of the results, including a discussion of the models' performance and their implications for different business applications. Finally, Section 6 concludes the report by summarizing key findings, highlighting contributions, and suggesting potential future research directions.

## 2. Literature Review

Hate speech detection has emerged as a critical area of research in Natural Language Processing (NLP) due to the escalating prevalence and harmful impacts of online abuse. This section provides an overview of established techniques, highlighting the evolution from traditional machine learning approaches to the recent breakthroughs achieved with large language models.

### 2.1. Traditional Machine Learning for Hate Speech Detection

Early research primarily relied on traditional machine learning algorithms for hate speech classification. These methods, including Naïve Bayes, Support Vector Machines (SVMs), and Logistic Regression, depend on the extraction of relevant features from text, such as:

* **Bag-of-Words (BoW):** Representing text as the frequency of individual words, disregarding grammar and word order.
* **Term Frequency-Inverse Document Frequency (TF-IDF):** Weighing the importance of words based on their frequency in a document and their rarity across a corpus.
* **Linguistic Features:** Incorporating grammatical information, sentiment analysis, and lexical resources (e.g., lists of hate words).

While these methods demonstrated some success, they often struggled to capture the nuanced linguistic patterns and contextual information crucial for accurately identifying hate speech.

### 2.2. Recurrent Neural Networks (RNNs) and LSTMs

Recurrent Neural Networks (RNNs) marked a significant advancement in NLP by introducing the concept of "memory," allowing them to process sequential data like text while considering the order of words and phrases. Long Short-Term Memory networks (LSTMs), a specialized type of RNN, further enhanced this capability by effectively capturing long-range dependencies in text, crucial for understanding context in hate speech detection. LSTMs have demonstrated improved performance compared to traditional ML methods, as they can automatically learn relevant features from the text itself, reducing the need for extensive feature engineering.

### 2.3 Large Language Models (LLMs) for Hate Speech Detection

The advent of large language models (LLMs), such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), has revolutionized NLP, including the field of hate speech detection. Trained on massive text corpora, LLMs develop a rich understanding of language and can discern complex patterns, making them highly effective for various downstream tasks.

**Fine-tuning:** Pre-trained LLMs can be fine-tuned on labeled hate speech datasets to adapt their knowledge for accurate classification.

**Zero/Few-Shot Learning:** LLMs' extensive pre-training allows them to generalize well to new tasks with minimal labeled data, making them particularly valuable for quickly adapting to evolving hate speech patterns.

### 2.4 Summary of Cited Papers

The following papers highlight the progress in using LLMs for hate speech detection:

* **BERT for Hate Speech Detection: A Comparative Study:** This study compared BERT's performance against traditional ML methods, demonstrating BERT's superiority in capturing contextual information and achieving higher accuracy in hate speech classification.
* **Fine-tuning GPT-3 for Hate Speech Detection: A Comparative Analysis:** Researchers explored fine-tuning GPT-3 for hate speech detection, showcasing its ability to outperform previous models, particularly in understanding nuanced language and context.
* **XLNet for Hate Speech Detection: Improving Hate Speech Detection with a Generative Pre-trained Transformer:** This work investigated the effectiveness of XLNet, another LLM architecture, for hate speech detection. The study emphasized XLNet's capacity to model bidirectional contexts and generate more accurate predictions.

These studies collectively underscore the potential of LLMs as powerful tools for combating online hate speech due to their ability to process large amounts of data, learn intricate linguistic patterns, and generalize to new and evolving forms of hate speech.

## 3. Data Collection and Pre-processing

This project leverages two datasets to train and evaluate the performance of different hate speech detection models: the Toxic Comment Classification dataset and the HateXplain dataset. This section details the characteristics of each dataset and outlines the pre-processing steps undertaken to prepare the data for model training and evaluation.

### 3.1 Toxic Comment Classification Dataset

The Toxic Comment Classification dataset, sourced from Wikipedia's talk page edits, comprises 159,571 user comments annotated for six types of toxicity:

* **Toxic**
* **Severe Toxic**
* **Obscene**
* **Threat**
* **Insult**
* **Identity Hate**

This dataset provides a multi-label classification problem, as individual comments can belong to one or more toxicity categories.

**Pre-processing:**

1. **Text Cleaning:**
   * Removal of punctuation.
   * Conversion of text to lowercase.
   * Removal of non-alphanumeric characters.
2. **Tokenization:** Splitting the text into individual words (tokens).
3. **Stop Word Removal:** Eliminating common words (e.g., "the," "a," "is") that provide little semantic meaning using the NLTK English stop word list.
4. **Lemmatization:** Reducing words to their base or root form (e.g., "running" to "run") using the WordNetLemmatizer.
5. **TF-IDF Vectorization:** Converting the text data into numerical feature vectors using TF-IDF, considering the 10,000 most frequent words in the corpus.

### 3.2 HateXplain Dataset

The HateXplain dataset, specifically designed for hate speech research, contains 150,000 social media posts and comments. In this project, we focus on the binary classification task, distinguishing between **hate speech** and **non-hate speech**. While the HateXplain dataset provides detailed annotations for various categories of hate speech, we simplify the task to a binary classification problem.

**Pre-processing:**

1. **Label Conversion:** The original labels were converted to a binary format:
   * **1:** Hate speech
   * **0:** Non-hate speech
2. **Text Combination:** The "post\_tokens" field (which contains tokenized text) was combined into a single text string for each instance.

### 3.3 Data Split

Both datasets were divided into training and testing sets using an 80%-20% split ratio. This division allows us to train our models on a larger portion of the data while reserving a separate portion to evaluate their performance on unseen examples.

4. Model Development

This section describes the models used in this comparative study, detailing their architectures, training processes, and rationale for selection. The models span traditional machine learning algorithms, a deep learning approach, and large language models, enabling us to evaluate their effectiveness in hate speech detection.

### 4.1 Traditional Machine Learning Models

Two commonly used traditional machine learning algorithms were employed, utilizing the pre-processed TF-IDF features extracted from the Toxic Comment dataset:

* **Logistic Regression:** This linear model is a simple and computationally efficient choice for binary and multi-class classification tasks. Its interpretability is a strength, as the coefficients associated with each feature provide insights into their importance for classification.

**Potential Concerns:**

* + While Logistic Regression has shown promise in some hate speech detection scenarios, its linear nature might not always adequately capture the complex interactions between words and phrases that are essential for nuanced hate speech recognition.
* **Random Forest:** This ensemble learning technique involves constructing multiple decision trees during training, then averaging their predictions to create a more robust model. Its robustness to overfitting is often cited as a major advantage.

**Potential Concerns:**

* + Random Forest, despite its strength, can be less interpretable than other models, making it difficult to understand the factors influencing its classification decisions. This lack of transparency can be problematic for certain applications where bias detection is crucial.

**Strengths of Traditional ML in this Context:**

* They offer a valuable baseline for comparison with the more complex deep learning and large language models.
* Their relative simplicity and computational efficiency make them well-suited for situations where resources or time are limited.

**Limitations of Traditional ML in this Context:**

* Their performance often depends on the quality of hand-engineered features, requiring careful feature engineering expertise.
* They might not be as adept as newer architectures (like LSTM and LLMs) in learning intricate patterns within complex text data, especially those related to nuances in hate speech.

### 4.2 Deep Learning Model: Long Short-Term Memory (LSTM)

To capture the dynamic and sequential nature of hate speech, we employed an LSTM network, a type of recurrent neural network (RNN) known for its effectiveness in processing sequential data, including text. The LSTM architecture we utilized comprised the following layers:

1. **Embedding Layer:** Embeds input words into dense vectors using pre-trained GloVe embeddings, capturing rich word semantics and facilitating the understanding of contextual information. By choosing not to train the embedding layer (trainable=False), we preserved the pre-trained word representations and allowed the model to leverage existing knowledge of word relationships.
2. **Spatial Dropout:** This regularization technique, applied to the embedding layer, randomly drops out input units to mitigate the risk of overfitting by reducing reliance on individual features and promoting a more robust model.
3. **LSTM Layer 1:** Processes the sequence of word embeddings with 128 units, enabling the model to learn and retain long-term dependencies within the input text.
4. **Dropout Layer:** An additional layer of regularization, applied after the first LSTM, further mitigates overfitting by randomly dropping out some connections between layers, further enhancing model generalization.
5. **LSTM Layer 2:** Continues processing the outputs of the first LSTM layer, decreasing the dimensionality of the representation to 64 units.
6. **Dense Layer:** The final output layer comprises 6 units (matching the number of toxicity categories in the Toxic Comment dataset), activated by the sigmoid function for multi-label classification.

**Training Details:**

* **Epochs:** 10
* **Batch Size:** 128
* **Optimizer:** Adam optimizer
* **Loss Function:** Binary Cross-Entropy (suitable for multi-label classification)

**LSTM Strengths in This Context:**

* Its ability to process sequential information allows it to capture dependencies and contextual relationships within text.
* Automatic feature learning through its internal architecture makes it less reliant on hand-crafted features compared to traditional ML models.

**Potential Concerns with LSTM:**

* Its high number of parameters can lead to longer training times and potential for overfitting if proper regularization strategies aren't in place.
* While GloVe embeddings are beneficial, they might not be ideal for the specific domain of hate speech.
* Deep learning models can sometimes struggle with interpreting and explaining their decisions, making it challenging to understand biases or inaccuracies in their outputs.

### 4.3 Large Language Models (LLMs)

We experimented with LLMs for hate speech classification on the HateXplain dataset:

* **BERT (Bidirectional Encoder Representations from Transformers):** A powerful transformer-based model pre-trained on a massive text corpus, learning intricate language patterns that allow it to discern complex semantic relationships and contextual understanding.
* **GPT-3 (Generative Pre-trained Transformer 3):** This large language model stands out for its exceptional ability to generate high-quality text, demonstrating advanced comprehension of natural language patterns.
* **GPT-2 (Generative Pre-trained Transformer 2):** A predecessor to GPT-3, also trained on a massive text corpus, possessing sophisticated language understanding capabilities.

**Experimentation Limitations:**

* The limited scale of LLM experiments was mainly due to the significant computational resources required for fine-tuning and running inference, coupled with the limitations of the OpenAI API (costs and rate limits).
* These factors restricted the amount of training and testing conducted, requiring a more conservative interpretation of the accuracy metrics.

**LLMs in the Context of Hate Speech Detection:**

* Their pre-training on massive text corpora equips them with a broader understanding of language compared to traditional ML or RNN-based models, making them potentially better suited for capturing the nuances of hate speech.
* Zero/few-shot learning approaches leverage their general language knowledge to predict with limited additional training, adapting quickly to evolving patterns of hate speech.

## 5. Results and Discussion

This section analyzes the performance of various hate speech detection models evaluated in this project. We'll examine their strengths and weaknesses based on key metrics, emphasizing their potential implications for different business applications.

### 5.1. Model Performance Overview

The table below summarizes the performance of each model on the Toxic Comment Classification dataset (multi-class) and the HateXplain dataset (binary classification for hate speech vs. non-hate speech):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Dataset** | **Category** | **Accuracy** | **Recall** |
| **Logistic Regression** | Toxic  Comment | Toxic | 0.91 | 0.73 |
| Severe Toxic | 0.54 | 0.33 |
| Obscene | 0.92 | 0.76 |
| Threat | 0.80 | 0.19 |
| Insult | 0.82 | 0.64 |
| Identity Hate | 0.71 | 0.32 |
|  |  |  |  |  |
| **Random Forest** | Toxic  Comment | Toxic | 0.87 | 0.74 |
| Severe Toxic | 0.43 | 0.14 |
| Obscene | 0.87 | 0.79 |
| Threat | 0.88 | 0.57 |
| Insult | 0.75 | 0.65 |
| Identity Hate | 0.61 | 0.17 |
|  |  |  |  |  |
| **LSTM** | Toxic  Comment | Toxic | 0.82 | 0.74 |
| Severe Toxic | 0.55 | 0.15 |
| Obscene | 0.83 | 0.79 |
| Threat | 0.12 | 0.57 |
| Insult | 0.73 | 0.65 |
| Identity Hate | 0.25 | 0.17 |
|  |  |  |  |  |
| **BERT** | HateXplain | Hate Speech (Binary) | 0.72 | N/A |
| **GPT-3** | HateXplain | Hate Speech (Binary) | 0.38 | N/A |
| **GPT-2** | HateXplain | Hate Speech (Binary) | 0.28 | N/A |

**Note:** Accuracy was the primary metric used to evaluate the LLMs (BERT, GPT-3, GPT-2) on the HateXplain dataset. The other metrics (precision, recall) for these models are not reported here due to potential inconsistencies in their calculation.

### 5.2 Analysis and Business Applications

**General Observations:**

* **Traditional ML (Logistic Regression, Random Forest):** Demonstrated relatively strong performance on the Toxic Comment dataset, particularly in achieving high precision for many categories. However, they often struggled with lower recall for categories like "Threat" and "Identity Hate," potentially indicating a limitation in detecting nuanced instances of hate speech.
* **LSTM:** Achieved comparable results to traditional ML models on Toxic Comment, but its performance varied across categories. Further analysis is needed to determine if its architecture effectively leverages the sequential nature of text for improved hate speech detection.
* **LLMs (BERT, GPT-3, GPT-2):** Experiments with LLMs on the HateXplain dataset yielded varied results, with BERT achieving the highest accuracy (72%) compared to GPT-3 (38%) and GPT-2 (28%). However, it's crucial to note that these results are based on limited experimentation due to the computational costs and rate limits associated with using the OpenAI API for fine-tuning and inference.

**Business Application Focus:**

* **Social Media Comment Moderation (Prioritize Recall):** Random Forest, with its generally higher recall on the Toxic Comment dataset, might be a good initial choice to minimize the risk of missing hateful content. However, its performance should be thoroughly evaluated on a dataset representative of the specific social media platform.
* **Brand Reputation Management (Prioritize Precision):** Logistic Regression, with its emphasis on high precision, could be more appropriate in scenarios where minimizing false positives and protecting brand image are paramount.
* **Content Filtering for Children's Platforms (Prioritize Both):** While BERT showed promising accuracy on HateXplain, further investigation and potentially a multi-layered approach combining automated detection with human review are essential for such sensitive applications.

**Additional Considerations:**

* **Limited LLM Experiments:** The LLM results should be interpreted with caution due to the limited experimentation. More extensive fine-tuning and evaluation are needed for a robust comparison with traditional ML approaches.
* **Dataset Biases:** Both datasets used in this project may contain biases that limit the models' generalizability to other contexts and demographic groups.
* **Explainability and Fairness:** As hate speech detection becomes increasingly reliant on sophisticated models, it's crucial to prioritize model explainability and fairness to ensure ethical and unbiased outcomes.

## 6. Conclusion

This project has undertaken a comparative analysis of diverse NLP techniques for hate speech detection, examining their efficacy across two datasets: the Toxic Comment Classification dataset and the HateXplain dataset. Our investigation has highlighted both the progress made in automated hate speech detection and the persistent challenges that need to be addressed for its successful application in real-world settings.

Key Takeaways:

* **Limitations of Traditional ML:** While traditional ML models achieved commendable overall performance in identifying toxic language, they exhibited limitations in detecting specific categories like "Threats" and "Identity Hate," suggesting their struggle to grasp the more subtle and contextualized forms of hate speech.
* **Promise of Deep Learning (LSTM):** The LSTM model showcased potential for capturing sequential information within text and exhibited strong performance on overall toxicity detection. However, its difficulties in capturing nuanced linguistic cues in certain categories, specifically related to threats and identity-based hate speech, mirror the challenges observed with traditional ML models.
* **Potential of Large Language Models (LLMs):** The preliminary results from our experimentation with BERT indicated significant potential for LLMs in understanding and detecting diverse hate speech forms, showcasing a greater capacity to analyze complex contextual information.

This research makes the following contributions:

* **Comprehensive Comparison:** This project provides a detailed comparison of different NLP techniques for hate speech detection, comparing the efficacy of traditional ML, deep learning (LSTM), and large language models.
* **Business Application Focus:** The analysis examines the practical implications of choosing models with varying precision and recall in specific business scenarios, such as social media moderation, brand reputation management, and content filtering for children's platforms.

Moving forward, future research could address several promising directions:

* **Advancing LLM Capabilities:** Investigating the potential of newer, more powerful LLM architectures like GPT-4 or incorporating advanced fine-tuning techniques for targeted improvements in hate speech detection is a key avenue for further exploration.
* **Ethical and Social Implications:** Developing explainable and interpretable hate speech detection models while actively addressing concerns related to fairness and potential bias across demographic groups is critical to ensuring ethical and responsible use of these technologies.

Ultimately, effectively combating online hate speech demands a holistic approach encompassing robust NLP techniques, a deeper understanding of the nuances of hateful language, and a commitment to ethical considerations.

## References

Almerekhi, H., Kwak, H., Salminen, J., & Jansen, B. J. (2020). Are these comments triggering? predicting triggers of toxicity in online discussions. In *Proceedings of the ACM Web Conference*, 2020, pp. 3366423. <https://doi.org/10.1145/3366423.3380074>

Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Alzamzami, F., & El Saddik, A. (2021). Monitoring cyber SentiHate social behavior during COVID-19 pandemic in North America. *IEEE Access*, *9*, 91184–91208.

Awal, M. R., Lee, R. K. W., Tanwar, E., Garg, T., & Chakraborty, T. (2023). Model-agnostic meta-learning for multilingual hate speech detection. *arXiv preprint arXiv:2303.02513*.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, (pp. 759–760).

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). *A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity*.

Baruah, A., Das, K. A., Barbhuiya, F. A., Dey, K. (2021). *IIITG-ADBU@HASOC-DRAVIDIAN-CODEMIX-FIRE2020: Offensive Content Detection in Code-Mixed Dravidian Text*. arXiv preprint arXiv:2107.14336.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., Sanguinetti, M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, (pp. 54–63). Association for Computational Linguistics. <https://doi.org/10.18653/v1/S19-2007>

Baydogan, C., & Alatas, B. (2021). Metaheuristic ant lion and moth flame optimization-based novel approach for automatic detection of hate speech in online social networks. *IEEE Access*, *9*, 110047–110062.

Bilal, M., Khan, A., Jan, S., & Musa, S. (2022). Context-aware deep learning model for detection of roman urdu hate speech on social media platform. *IEEE Access*, *10*, 121133–121151.

Boishakhi, F. T., Shill, P. C., & Alam, M. G. R. (2021). Multi-modal hate speech detection using machine learning. In *2021 14th International Conference on Electrical and Computer Engineering (ICECE)*, (pp. 4496–4499).

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, (pp. 491–500).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., ... & Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 33, 1877-1887.

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, *7*(2), 223–242.

Burnap, P., & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*, 1–15.

Caselli, T., Basile, V., Mitrovi"c, J., Kartoziya, I., & Granitzer, M. (2020). *I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language (pp. 6193–6202)*. Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, Y., & Pan, F. (2022). Multimodal detection of hateful memes by applying a vision-language pre-training model. *PLoS One*, *17*(9), e0274300.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(2), 71-80.

Chiril, P., Pamungkas, E. W., Benamara, F., Moriceau, V., & Patti, V. (2022). Emotionally informed hate speech detection: A multi-target perspective. *Cognitive Computation*, *14*, 322–352.

Chiu, K.-L., Collins, A., & Alexander, R. (2022). *Detecting hate speech with GPT-3*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., ... & Conneau, A. (2020). *Unsupervised cross-lingual representation learning at scale*.

Curry, A. C., Abercrombie, G., & Rieser, V. (2021). *ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI*. *arXiv preprint arXiv:2109.09483*.

Das, M., Saha, P., Dutt, R., Goyal, P., Mukherjee, A., & Mathew, B. (2021). You too Brutus! trapping hateful users in social media: Challenges, solutions & insights. *arXiv preprint arXiv:2108.00524*.

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, (pp. 86–95).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(3), 1–30.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In \*Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics