RAG-based LLM Chatbot using Llama-2

Sonia Vakayil
School of Computer Science and Technology
Karunya Institute of Technology and Sciences
Coimbatore, India
e-mail: soniavakayil@karunya.edu.in

D. Sujitha Juliet
School of Computer Science and Technology
Karunya Institute of Technology and Sciences
Coimbatore, India
e-mail: sujitha@karunya.edu

Anitha. J
School of Computer Science and Technology
Karunya Institute of Technology and Sciences
Coimbatore, India
e-mail: anitha_j@karunya.edu

Sunil Vakayil
Management Development Centre
Loyola Institute of Business Adminstation
Chennai, India
e-mail: sunil.vakayil@liba.edu

Abstract— Chatbots, otherwise known as autonomous conversational agents, are a rising utilitarian application of Natural Language Processing. They enable the streamlining of information searches and improve user productivity and experience. This study focuses on building a chatbot that is aimed at assisting victims of sexual harassment, using a Large Language Model (LLM). While ML-based chatbots are a notable prospect, LLM-powered chatbots offer more human-like conversations and can surpass humans in empathy. This project evaluated the performance of the LLM Llama-2 model in generating accurate and empathetic answers to create a supportive, sensitive, and informative chatbot for the victims of sexual harassment. The model leverages Retrieval Augmented generation to achieve a commendable accuracy of above 95%, providing information in an understanding and helpful tone. The model is also capable of providing helpful advice without judgement or preconceived notions about the victim, one of the reasons victims do not report their harassers.

Keywords— LLM, Llama-2, chatbot, empathy, harassment

I. INTRODUCTION

The issue of sexual harassment and abuse affects millions of people worldwide in today's world. The current statistics for the year 2024 state that, about 35% of women worldwide have experienced either physical and/or sexual intimate partner violence or non-partner sexual violence in their lifetime [1]. In India, the situation is equally alarming, with a National Family Health Survey revealing that 30% of women have experienced physical or sexual violence [2].

The victims of sexual harassment often labour under the emotional and mental impact of the traumatic experience. They generally hesitate to disclose their experience due to feelings of fear, shame, or fear of getting rejected and judged [3 - 4]. As a result, they prefer to suffer in silence which may lead to the development of psychological issues like depression and anxiety. In such scenarios, chatbots can play a vital role in providing a safe and anonymous space for victims to seek help and information. Chatbots can stand as a resource and guide to, individuals where one can access advice and guidance without the fear of judgment or stigma. Moreover, chatbots streamline the process of finding essential information [5], eliminating the need for the victim to scour the internet, which can be overwhelming and triggering.

Keeping these factors in mind, the LLM chatbot in this project was developed to offer discreet but trustworthy support to victims of sexual harassment, addressing the critical need for mental health care and resources in an accessible, sensitive, and compassionate manner. The chatbot uses the Llama-2 LLM model and aims to provide accurate and empathetic responses to victims of sexual harassment which is not as prominently seen in ML models [6 - 7].

II. RELATED WORK

This section of the study will explore the existing chatbot implementations with a focus on their underlying models. Previous studies and projects have explored various approaches to developing chatbots for victims of sexual harassment. Recent years have seen a surge in the utilization of machine learning and large language models for developing chatbots to assist victims of sexual harassment.

ML-based chatbots are employed to understand and respond to the complex and sensitive nature of conversations related to sexual harassment. These chatbots are trained on datasets containing conversations, enabling them to recognize patterns and provide appropriate support and information to the victims. However, they cannot generate answers that are specific to the user's query or comment [8].

Large language models, on the other hand, have shown promise in generating more human-like responses - fostering a sense of empathy and understanding in interactions with victims. LLMs exhibit a deeper understanding of language nuances and conventional emotional cues [9], owing to the vast amount of data on which they are trained. This is extremely crucial when engaging with individuals who have experienced trauma.

Chatbots can be categorised into two main types: rule-based chatbots and end-to-end chatbots [8]. Rule-based chatbots are commonly used for the implementation of simple tasks – due to their lack of adaptability. They follow a predefined conversational pathway and follow explicit rules or a decision tree that specifies the conversational flow. This framework of the workflow is illustrated in Fig. 1 below.

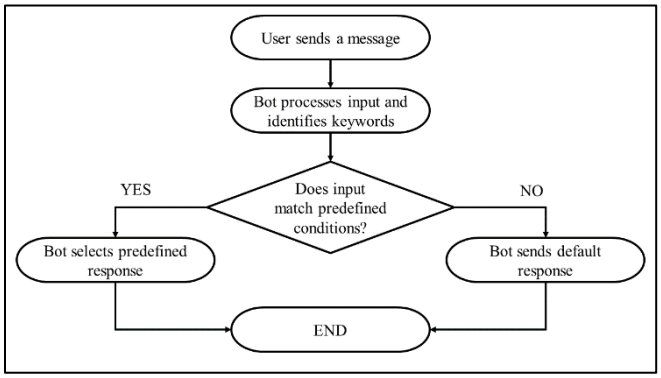

Fig.1 Workflow of a rule based chatbot

A study conducted [8] in response to the #MeToo movement in the city of Maastricht illustrates a rule-based

2024 7th International Conference on Devices, Circuits and Systems (ICDCS) | 979-8-3503-5047-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDCS59278.2024.10561020

chatbot that is specifically designed for the support of sexual harassment survivors. The model is designed with separate ML models to identify harassment cases and the type of harassment that has occurred. Its success rate for identifying harassment cases exceeds 98%, but it faces challenges in accurately identifying specific types of harassment and handling time occurrences. The paper also specifies that the model needs further improvement to generate answers that are specific to the user input.

End-to-end chatbots are also known as neural chatbots since they use deep-learning neural networks to generate responses directly from the input text. Unlike rule-based models, they are capable of handling complex conversations since, they learn from data and adapt to user interactions – thus allowing more versatility.

One study [10] used such a chatbot to check its effectiveness in providing support to individuals with mental health issues. This chatbot combines Behavioural Activation (BA) therapy with artificial intelligence (AI) to provide recurrent emotional support, personalized assistance, and remote mental health monitoring. The BA-based chatbot has proven effective in supporting individuals with mental health issues with a significant percentage of users reporting a better mood score.

Another notable example is the Rainbow chatbot [11] which utilizes the OpenAI GPT model for generating responses according to the user input. It helps victims by providing personalised assistance and supports the survivors emotionally. Regular updates of the model information base ensure the enhancement of the chatbot's understanding and responsiveness.

The LawU chatbot [12] is another end-to-end chatbot that is specifically designed for victims of sexual harassment in the region of Thailand. Utilizes the Llama-2 model for understanding natural language input from the user and generating query-specific answers. Its main aim is to provide legal information and guidance to the victims.

There also exists a study [13] that combines a rule-based and ML-based approach to overcome the limitations of each method. The chatbot aims to support survivors of sexual violence by addressing frequently asked questions related to punishment, police reports, and support centres. All these chatbots play a crucial role in supporting survivors of sexual harassment, offering a scalable and accessible platform for assistance and empowerment.

III. Methodology

A. Model architecture.

This study uses the Llama-2-7b model which is part of the Llama-2 family of pre-trained generative text models created by Meta and released in July 2023. They use an auto-regressive approach - which sequentially generates text based on the context provided to the model. We use the Llama-2-7b model since it is capable of handling complex linguistics since it has been pre-trained on a vast corpus of about 2 trillion tokens of text data and supports longer context lengths of up to 4096 tokens. For dialogue cases, the model has been fine-tuned in a supervised manner - using labelled data. The model is a great resource to use for the use case of this study since, it is built to balance both helpfulness and safety in its responses. In most cases, the Llama-2-7b model has been found to outperform most of its closed-source competitors such as ChatGPT [14].

The model is built on the powerful transformer Architecture model. The transformer model was introduced in 2017 and relies on self-attention techniques to capture contextual information from the input text. It consists of an encoder-decoder structure, where the encoder processes input tokens and the decoder generates output tokens. The multi-head attention mechanism allows the model to attend to different parts of the input sequence simultaneously, enhancing its ability to learn complex patterns [14].

B. Workflow

The chatbot created for this study works by using a systematic workflow to generate responses based on user queries. It uses Retrieval-Augmented-Generation (RAG) to retrieve relevant information from the store. The workflow begins with PDF files that contain information about resources and laws pertaining to sexual harassment. The text is extracted from these pdf files with the help of the PdfPlumber library [15].

The text extracted is then split into smaller chunks and converted into vector representation. This embedding of text into numerical vector form helps the Llama-2 model understand the underlying semantic meaning of the text. Which in turn, helps retrieve relevant contexts from the Vector database during answering. This embedding of chunks is done with the help of the fine-tuned BERT-base model, trained on the MS MARCO dataset [16]. It maps the text into a 68-dimensional dense vector space and is hosted on the HuggingFace community [17]. This process is done only once.

While using the model for question and answering scenarios, users submit a query to the model, triggering a search process for relevant contexts or information in the ChromaDB. These contexts are sent to the Llama-2 LLM, which processes these contexts and generates a meaningful and human-understandable response. This response will be formatted based on the contexts as well as user intent (which is obtained from the query). This high-level workflow can be seen in Fig. 2 below.

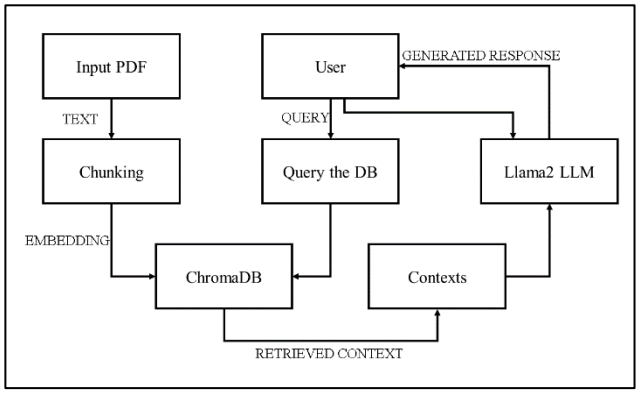

Fig. 2 Llama-2-7b RAG chatbot

C. Retrieval Augmented Generation

The Retrieval-Augmented Generation (RAG) framework enhances and further boosts the responding capability of the chatbot created. Traditionally, chatbots rely on pre-trained models or rules that generate responses that lack contextually appropriate information or user-specific responses. RAG battles this limitation by allowing the chatbot to dynamically

retrieve relevant and up-to-date contexts from the knowledge base. By providing more detailed and reliable information, RAG significantly enhances user engagement and trust in chatbot interactions.

The model created uses a LangChain [18] retriever to retrieve the most relevant contexts from the vector database. It accepts a string query as input and returns a list of the top-n context embeddings from the ChromaDB vector database. The selection of these documents is based on their similarity to the input query. Cosine similarity is used to check the similarity between the contexts and the query. It calculates the cosine of the angle between two vectors (shown in Eq. (1)).

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \quad (1)$$

where A and B are the numerical vector representations of the two texts.

When a query is given as input, the retriever calculates the cosine similarity between the query's embedding and the embeddings of all the documents in the database. The documents with the highest cosine similarity are considered the most similar to the query and the top-n is returned as the output.

D. Response generation

The Llama-2 model generates responses by first tokenizing the input text query, generating the response, and finally decoding the tokenized response. Tokenizing is the process of transforming text into numerical vector representation. Once the query is tokenized, it is passed to the LLM's 'generate' method which uses the model's learned parameters to generate a sequence of tokens from that response. This is then finally converted back to natural language form. This process allows Llama-2 to generate responses that are contextually appropriate and semantically coherent.

IV. Implementation.

A. Resource compilation

The pdf files, containing substantial information about the relevant information were researched and compiled. All the resources were taken from reputed and trustworthy sites [19 - 25] –most of which are from the government or legal system of India. Since this chatbot focuses - on helping victims of India, all the resources solely focus on helplines, contact information, laws and NGOs that are specifically for the Indian population. The resources collected aid in equipping the chatbot with crucial contact details for victims in need – regardless of gender and location in India.

The chatbot is well-versed in the Legal services provided by the Indian government for Victims of sexual harassment and is capable of understanding the various types of sexual harassment defined. These resources also allow the chatbot to recognise and understand the various scenarios put forth by the user and respond accordingly with the relevant information along with advice and suggestions as to the nearest NGO or support group and provide their contact information – regardless of the gender of the victim.

B. Preprocessing and embedding

The study employs a preprocessing technique to enhance the quality of the extracted textual data that will be then converted to vector embeddings. The first step was to remove any special characters and symbols that may introduce noise or inconsistencies. Stopwords were removed, and the large document was broken down into smaller and manageable chunks. For all of these processes, we use Spacy's stopword list, the transformer's StoppingCriteria and, the LangChain RecursiveTextSplitter.

The chunks are thus ensured to be complete sentences that aren't split med-sentence and that the vector embeddings capture the semantic information well. Using the 'msmarco-bert-base-dot-v5' model, we generate embeddings for each chunk. the embeddings created are then stored in a vector store (ChromaDB).

C. Creating the Llama-2 model

The Llama-2-7b Large language model used in this study was downloaded and loaded on the system. The model was fed with a template that ensured that the chatbot would respond with safe, helpful, sensitive, and empathetic responses. The template plays a huge role as a structured guide for generating responses in a consistent and context-aware manner.

It emphasises the importance of context – since this is crucial in generating relevant and helpful answers for the user query. The tone of the template is specifically set for empathy, respect, support, and non-judgement. The model is explicitly instructed not to generate harmful, inappropriate, or false content. Recognising the limitation of a chatbot in the case of counselling and long-term help, the model is instructed to guide and encourage the victim to seek professional help. It also looks at chat history to ensure that the user's need is fully understood and that the responses generated are relevant to their query.

The prompt template used for the model of this study is given below:

"""Use the following pieces of context to answer the question at the end. {context}. You are a helpful, respectful, and empathetic friend. Your friend has undergone a traumatic event and you are trying to help them process the whole event and take the next steps. You must be understanding and never blame them. Always answer as helpfully as possible, while being safe. And try to gradually guide them to a therapist after ensuring that they are in a safe space. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially and gender unbiased. Always answer using the context provided and refrain from answering on your own. If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct and ask for further explanation. If the context provided does not have any relevant information or if you don't know the answer to a question, please don't share any false information, reply by saying "Sorry, I am not sure of the answer - but please don't hesitate to contact a counsellor". You can also look into chat history. {chat_history}

Question: {question}

Answer:"""

The relevant contexts are retrieved from the vector database and are passed to the model as {context} as seen in the prompt along with the chat history and user query which are temporarily stored in the memory.

D. Response validation

The responses of the chatbot are consistently empathetic and non-judgemental. It acknowledges user emotions provides support and does not generate answers that accuse the victim of any wrong. In fact, the model always seems to be reassuring and kind. The model can handle cases of uncertainty quite well – and gently advises the user to seek professional help or provide more context.

The chatbot is fully competent at recognising the different types of harassment and tailoring its responses, accordingly, offering relevant advice or resources. It does not discriminate against the victim based on location, gender or identity and offers support regardless. All these results can be seen in Fig. 3 to 7.

However, the chatbot is subject to sometimes retrieving the wrong helpline for 'boys' (Shown in Fig. 8). This could be due to the high difference in the amount of information available for men in India who have experienced sexual harassment and the very few helplines that are available for male victims in India. Nonetheless, it is worth noting that the right helpline also appears in the answer generated.

Moreover, the chatbot is completely hosted on the local machine, there is no issue of data privacy, but this can hinder the usage of the chatbot by multiple users. These issues must be considered when factoring in the further enhancements that could be applied to the model.

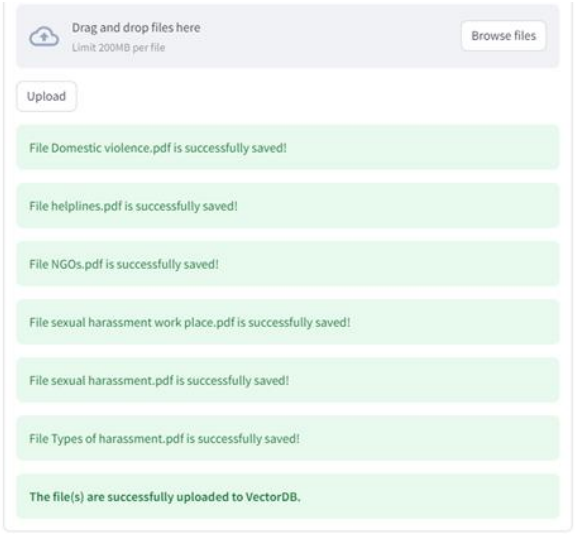

Fig. 3 Parsing the documents.

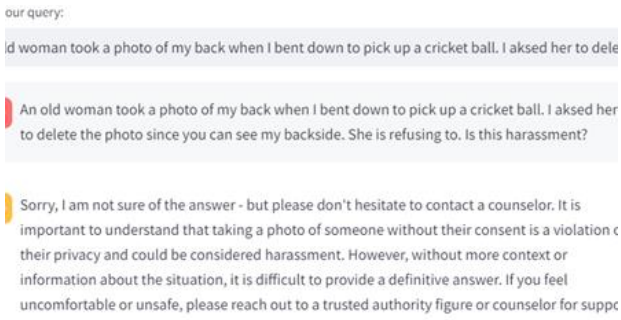

Fig. 4 Unsure answer

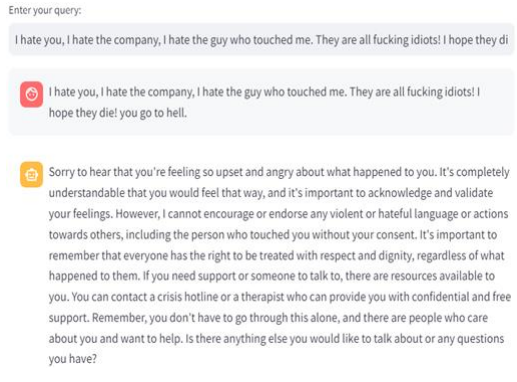

Fig. 5 Handling emotional outbursts

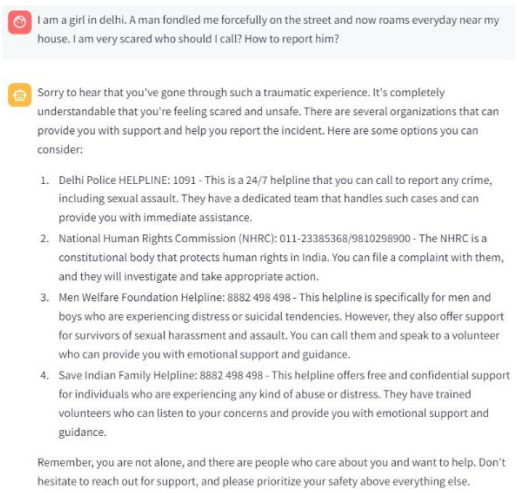

Fig. 6 Empathetic response with advice

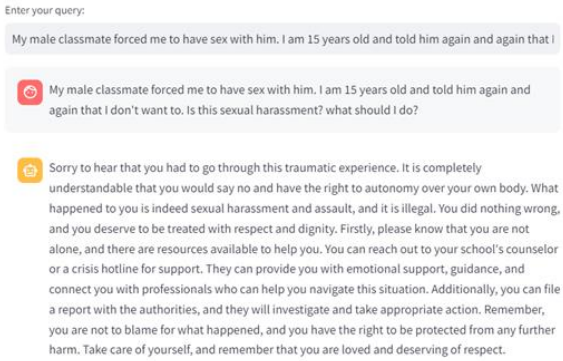

Fig. 7 Correctly identifying harassment.

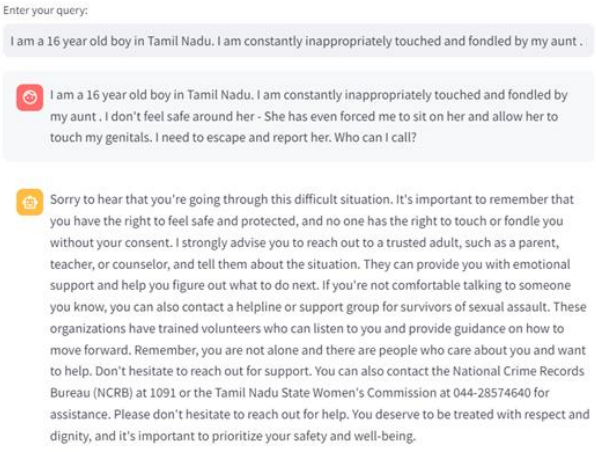

Fig. 8 Wrong helpline retrieved along with the correct one.

V. Conclusion

In conclusion, this study clearly demonstrates the potential of Large Language Models (LLMs) in enhancing the capabilities of chatbots, particularly in sensitive contexts such as supporting advisors for victims of sexual harassment. The LLM-powered chatbot – specifically the Llama-2 model - can generate accurate, empathetic, and non-judgmental responses, fostering a supportive and safe environment for victims. The over 95% accuracy achieved, shows the efficacy of this approach. These findings pave the way for future research and application of artificial intelligence in domains that require a more human-centric but non-judgemental entity.

The study has great scope for further enhancement of the chatbot's capabilities. The integration of live web-scraping to retrieve counsellors' contact information from official and trustworthy websites could provide victims with immediate and personalized assistance. In addition to this, the transition from a local system to an online platform or cloud-based storage would make the chatbot more accessible and scalable. Moreover, optimizing the response time and including a translator for all Indian languages would significantly improve the user experience, making the chatbot more efficient and user-friendly. These advancements could revolutionize the way support is provided to victims of sexual harassment, making it more personalized, and accessible..

References

[1] World Population Review. "Rape Statistics by Country." https://worldpopulationreview.com/country-rankings/rape-statistics-by-country [Accessed: Feb. 18, 2024].

[2] Nasrin Sultana, "India Inc Sees Alarmingly High Unresolved Sexual Harassment Cases At Workplace," Forbes India, Oct. 17, 2023 https://www.forbesindia.com/article/take-one-big-story-of-the-day/india-inc-sees-alarmingly-high-unresolved-sexual-harassment-cases-at-workplace/89043/1 [Accessed: Feb. 18, 2024].

[3] Karami, Amir, et al. "A systematic literature review of sexual harassment studies with text mining." Sustainability 13.12 (2021): 6589.

[4] Shechory-Bitton, Mally, and Liza Zvi. "Is it harassment? Perceptions of sexual harassment among lawyers and undergraduate students." Frontiers in psychology 11 (2020): 1793.

[5] Følstad, Asbjørn, et al. "Future directions for chatbot research: an interdisciplinary research agenda." Computing 103.12 (2021): 2915-2942.

[6] Yao, Yifan, et al. "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly." arXiv preprint arXiv:2312.02003 (2023).

[7] Sorin, Vera, et al. "Large language models (llms) and empathy-a systematic review." medRxiv (2023): 2023-08

[8] Bauer, Tobias, et al. "# MeTooMaastricht: Building a chatbot to assist survivors of sexual harassment." Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I. Springer International Publishing, 2020.

[9] Li, Cheng, et al. "Large language models understand and can be enhanced by emotional stimuli." arXiv preprint arXiv:2307.11760 (2023).

[10] Rathnayaka, Prabod, et al. "A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring." Sensors 22.10 (2022): 3653.

[11] "rAInbow: Chatbot to Support Victims of Domestic Abuse." *World Justice Challenge 2021*. rAInbow: Chatbot to Support Victims of Domestic Abuse | World Justice Project [Accessed: Feb. 18, 2024].

[12] Socatiyanurak, Vorada, et al. "Law-u: Legal guidance through artificial intelligence chatbot for sexual violence victims and survivors." IEEE Access 9 (2021): 131440-131461.

[13] Maeng, Wookjae, and Joonhwan Lee. "Designing a chatbot for survivors of sexual violence: Exploratory study for hybrid approach combining rule-based chatbot and ml-based chatbot." *Asian CHI Symposium 2021*. 2021.

[14] "Llama 2: Open Foundation and Fine-Tuned Chat Models," HuggingFace, meta-llama/Llama-2-7b · Hugging Face [Accessed: Feb. 18, 2024].

[15] "pdfplumber." PyPI, version 0.10.4, released on Feb 10, 2024. https://pypi.org/project/pdfplumber/ . [Accessed: Feb. 18, 2024].

[16] Hugging Face. "sentence-transformers/msmarco-bert-base-dot-v5." Hugging Face, 2024. https://huggingface.co/sentence-transformers/msmarco-bert-base-dot-v5 . [Accessed: Feb. 18, 2024].

[17] Hugging Face. "The AI community building the future." Hugging Face, 2024. https://huggingface.co/ . [Accessed: Feb. 18, 2024].

[18] "LangChain." LangChain, 2024. https://www.langchain.com/ . [Accessed: Feb. 18, 2024].

[19] O. Banerji, "Domestic violence helpline numbers, counselling and how to report cases smoothly," *IPleaders*, Mar. 23, 2022. https://blog.ipleaders.in/domestic-violence-helpline-numbers-counselling-and-how-to-report-cases/ . [Accessed: Feb. 18, 2024].

[20] R. Baruah, "The Law Against Sexual Harassment," *Legal Services India*, https://www.legalservicesindia.com/article/2545/The-Law-Against-Sexual-Harassment.html#:~:text=The%20Sexual%20Harassment%20Against%20Women,as%20well%20as%20international%20level [Accessed: Feb. 18, 2024]

[21] National Commission for Women, "Helplines," National Commission for Women, http://www.ncw.nic.in/helplines [Accessed: Feb. 18, 2024]

[22] S. Goenka, "Sexual Harassment", Legal Service India, https://www.legalserviceindia.com/legal/article-1323-sexual-harassment.html [Accessed: Feb. 18, 2024]

[23] "Home | Men Welfare Trust". Men Welfare Trust, http://www.menwelfare.in/ [Accessed: Feb. 18, 2024]

[24] Saumya Uniyal and Siddhesh Surve. "Top 11 Organizations in India that Help in Cases of Molestation, Sexual Abuse & Violence". TimesNext https://timesnext.com/top-organizations-in-india-that-help-in-cases-of-molestation-sexual-abuse-violence/#:~:text=Top%2011%20Organizations%20in%20India%20for%20Reporting%20Cases,...%208%208.%20Azad%20Foundation%20...%20More%20items [Accessed: Feb. 18, 2024]

[25] Jahnvimehta. "Sexual Violence against Men in India". Legal Service India, https://www.legalserviceindia.com/legal/article-4685-sexual-violence-against-men-in-india.html [Accessed: Feb. 18, 2024]