Decoding Concerns: Multi-label Classification of Vaccine Sentiments in Social Media

Somsubhra De[1,*,†], Shaurya Vats[2,†]

[1]Indian Institute of Technology Madras, Tamil Nadu, India
[2]Indian Institute of Technology Kharagpur, West Bengal, India

Abstract

In the realm of public health, vaccination stands as the cornerstone for mitigating disease risks and controlling their proliferation. The recent COVID-19 pandemic has highlighted how vaccines play a crucial role in keeping us safe. However the situation involves a mix of perspectives, with skepticism towards vaccines prevailing for various reasons such as political dynamics, apprehensions about side effects, and more. The paper addresses the challenge of comprehensively understanding and categorizing these diverse concerns expressed in the context of vaccination. Our focus is on developing a robust multi-label classifier capable of assigning specific concern labels to tweets based on the articulated apprehensions towards vaccines. To achieve this, we delve into the application of a diverse set of advanced natural language processing techniques and machine learning algorithms including transformer models like BERT, state of the art GPT 3.5, Classifier Chains & traditional methods like SVM, Random Forest, Naive Bayes. We see that the cutting-edge large language model outperforms all other methods in this context.

Keywords

Multi label classifier, Vaccine skepticism, Machine Learning, AI, COVID-19, Concerns, Sentiment, Tweet classification, LLM, Prompt engineering, Transformer models

1. Introduction

In the wake of the global COVID-19 pandemic, misinformation and vaccine hesitancy have emerged as significant societal challenges. The spread of anti-vaccine sentiment on social media platforms has amplified concerns. We address the complex task of multi-label classification of anti-vax tweets, aiming to identify and categorize the diverse spectrum of vaccine-related misinformation in the era of COVID-19. Through experimental evaluation, we demonstrate the efficacy of our proposed approach - our results not only underscore the prevalence of vaccine related concerns in social media discussions but also shed light on the nuanced nature of these concerns. The work provides valuable insights for policymakers, health organizations to communicate and intervene where required and emphasizes the significance of utilizing real-time social media discussions to inform evidence-based strategies that address public concerns

Forum for Information Retrieval Evaluation, December 15-18, 2023, Panjim, India

[*] Corresponding author.

[†] These authors contributed equally.

somsubhra@outlook.in (S. De); shauryavats56@gmail.com (S. Vats)
https://linkedin.com/in/somsubhrad (S. De); https://linkedin.com/in/shaurya-vats-6a521a231 (S. Vats)
0000-0003-3839-2575 (S. De); 0009-0006-0283-2064 (S. Vats)

CEUR Workshop Proceedings (CEUR-WS.org)

and promote informed decision-making.
All the materials, including code, datasets, and related resources for our work, are accessible in this repository.

2. Understanding the Task

Given a tweet content, it can be classified into categories from the list of the following 12 total classes (each of the classes refer to the different concerns or reasons against the use of vaccines) that we have:

- **Unnecessary** - The tweet indicates that vaccines are unnecessary or alternate cures are better
- **Mandatory** - The tweet is against mandatory vaccination
- **Pharma** - The tweet indicates that the Big Pharmaceutical companies are just trying to earn money, or it is against such companies in general because of their history
- **Conspiracy** - The tweet suggests some deeper conspiracy, and not just that the Big Pharma want to make money (e.g. vaccines are being used to track people, COVID is a hoax)
- **Political** - The tweet expresses concerns that the governments or politicians are pushing their own agenda though the vaccines
- **Country** - The tweet is against some vaccine because of the country where it was developed or manufactured
- **Rushed** - The tweet expresses concerns that the vaccines have not been tested properly or the published data is not accurate
- **Ingredients** - The tweet expresses concerns about the ingredients present in the vaccines (e.g. fetal cells, chemicals) or the technology used (e.g. mRNA vaccines can change the DNA)
- **Side-effect** - The tweet expresses concerns about the side effects of the vaccines, including deaths caused
- **Ineffective** - The tweet expresses concerns that the vaccines are not effective enough and are useless
- **Religious** - The tweet is against vaccines because of religious reasons
- **None** - No specific reason stated in the tweet or some reason other than the given ones

As this task involves multi-label classification, each tweet can be mapped to more than one labels, contingent on the stances expressed within the text. Below are a few tweet examples, each accompanied by associated labels and detailed descriptions to provide better context.

3. Dataset

The training dataset, comprising **9,921 anti-vaccine** tweet texts (posted during 2020-21) with corresponding tweet IDs and annotated labels, was sourced from the CAVES[1] Dataset. The test set consists of 486 tweets (along with their IDs), encompassing various vaccine types such as MMR, flu vaccines and more, in addition to COVID-19 vaccines.

Table 1
Examples of Tweet Classifications

Tweet Text	Labels	Tweet Description
@jeffmcnamee @Amanda77197114 @alexanderchee BREAKING: FDA announces 2 deaths of Pfizer vaccine trial participants from "serious adverse events.â€. Fed Up Democrats Say NO to Forced Vaccines in NY	side-effect mandatory political	*The tweet reports FDA's announcement of two deaths during Pfizer vaccine trials due to 'serious adverse events' and mentions that members of the Democratic party in New York are expressing their strong opposition to the idea of mandatory or compulsory vaccinations.*
My Take On The new vaccines? 1. I don't trust pharmaceutical companies 2. Only 61% of public will get vaccinated 3. Problems with people with allergies 4. They're not free 5. No data on longevity 6. Side effects? 7. Children under 16 were not tested in the Pfizer trial https://t.co/vZ4ZPkroc4	side-effect pharma rushed	*The tweet expresses skepticism about new vaccines, citing distrust of pharmaceutical companies, concerns about side effects of rushed vaccines, and other issues, particularly focusing on the lack of comprehensive testing.*
Why would we get a vaccine with so many side effects? Ccp covid has a 99.7% survival rate, so why get a vaccine? Thatâ€™s stupid and dangerous.	side-effect unnecessary	*The tweet questions the need for vaccines due to perceived side effects and argues against it, citing a high survival rate.*
@TorontoStar Then I suppose there is no hope for a vaccine. Nothing coming and it won't work if it does come. It is Lockdowns Forever, or take your chances with Covid. Lockdowns are no life at all. My choice it to take my chances. Bring it on!	ineffective	*The tweet expresses skepticism about the effectiveness of vaccines, suggesting that there's no hope for them and indicating a preference for taking their chances with COVID-19 over enduring continuous lockdowns.*
Oh my! One doesn't have to be an expert at reading body language to know he's covering something up. Depopulation is his game.	conspiracy	*The tweet implies a conspiracy claiming that someone is involved in an agenda, planning to reduce the global population intentionally.*

3.1. Some insights about Train set

Performing an EDA, we observe that the dataset exhibits class imbalance with certain labels having significantly higher counts compared to others. The two primary concerns `side-effect` and `ineffective` appear to dominate the dataset where as vaccine skepticism due to religious reasons and concerns related to the country of origin are less prevalent.

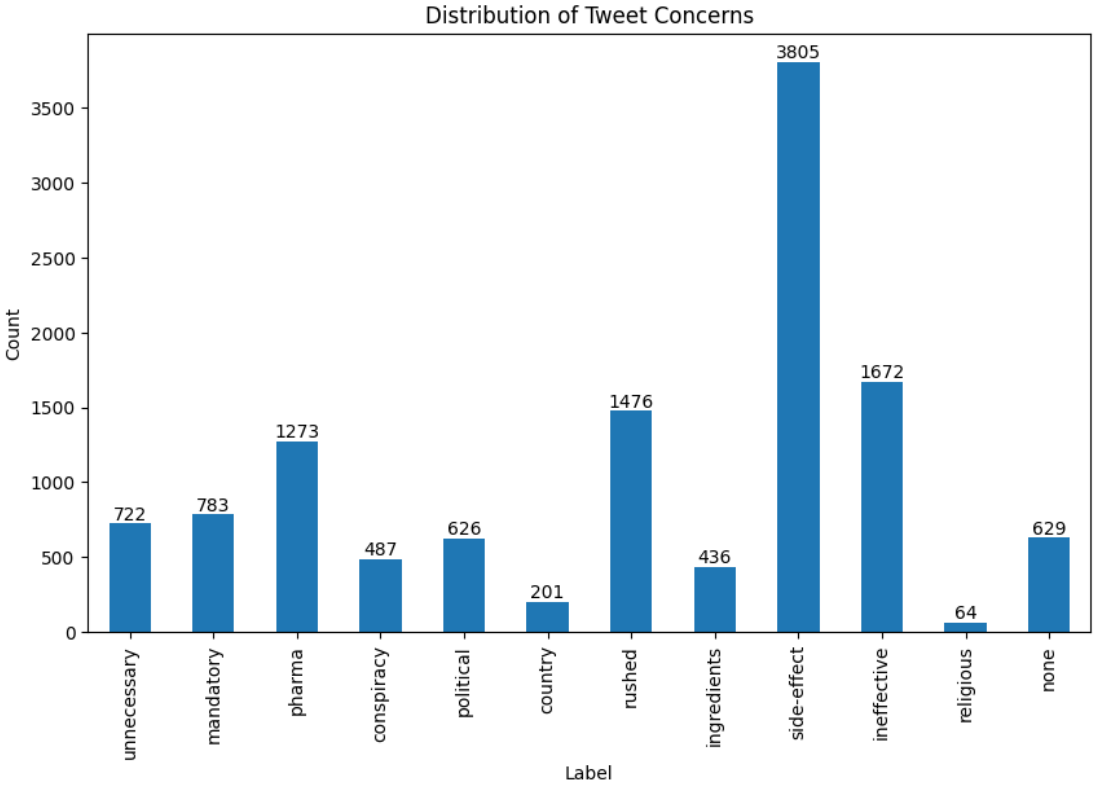

Figure 1: Categorical Data split (Train set)

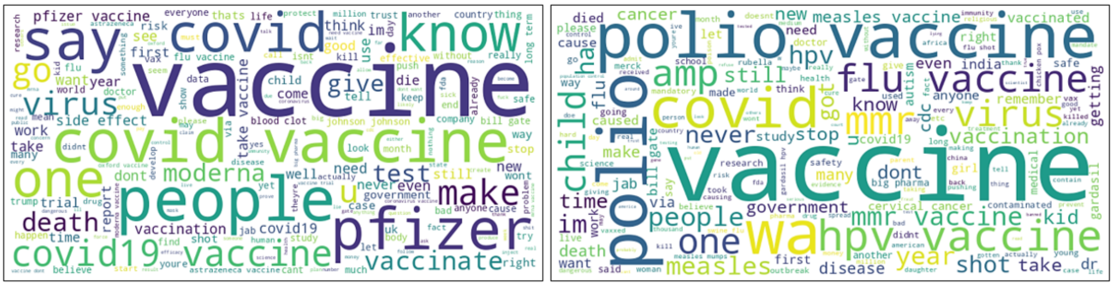

Figure 2: Word Cloud visualization of Train (left) vs. Test set Data (right)

4. Our Methodology

4.1. Data Pre-processing

To enhance the model's performance, we conducted comprehensive data pre-processing on both the training and test datasets. This involved meticulous cleaning to eliminate noisy data, and various other steps ensuring that the model's integrity remained intact.

- Stop word Removal: Elimination of commonly occurring English stop words (eg, ‘a’, ‘an’, ‘the’, etc. which do not add much meaning to the text) as per the NLTK library
- Lower casing Text: Conversion of the tweet text to lowercase, ensuring consistent, uniform handling of words
- Removal of punctuations and twitter handles or usernames (‘@’)
- Removal of non-alphanumeric characters excluding hashtags
- Emoji Translation: Converting emojis to their corresponding textual representations
- Contraction Expansion: Expansion of contractions to their full word forms, enhancing text clarity and comprehensibility.
- Removal of URLs as they don’t contribute to sentiment analysis

4.2. Tokenization & Lemmatization

Tokenization refers to the process of breaking down of a sequence of text into smaller units or tokens. It is crucial for feature extraction, dimensionality reduction, normalization and semantic understanding. The tweets have been tokenized using the `nltk.tokenize` package. **Lemmatization** is a text normalization technique, it is the process of reducing words to their meaningful base form (lemma) to ensure variants of a word are treated as a single item for analysis or retrieval purposes. We employed the `WordNetLemmatizer` from the NLTK library for this task. We chose lemmatization over an alternative technique called stemming for a specific reason. Stemming involves chopping off the ends of words to reduce them to a common root. While this can be useful in some cases, it’s not ideal when preserving the full meaning of words is important. Stemming can sometimes truncate words too aggressively, resulting in a loss of meaning. In our context, where understanding the context of text is crucial, lemmatization helps us retain more of the original word’s meaning while still achieving normalization. Additionally, we integrated **Part-of-Speech (PoS) tagging** into our text processing pipeline which involves assigning a grammatical category (nouns, verbs, adjectives, etc.) to each word in the text. This further enhances our ability to understand the context and helps us make more precise decisions during lemmatization.

5. Models

5.1. LLM

We have employed Large Language Models (LLMs) to address a particularly intricate and under-explored realm - multi-label classification, within the context of Social Media Tweets on vaccines. Our exploration involved the utilization of diverse prompting methods, such as chain of thought (CoT) style[2], zero-shot & few-shot learning. Through rigorous experimentation, we discovered that an amalgamation of these prompting styles proved most effective for our objectives. Subsequently, we formulated a novel prompt template tailored specifically for multi-label text classification tasks with LLMs. We noted an enhancement in model performance when prompts requiring the model to generate explanations or reasonings were incorporated. The rationale behind the model’s assignment of labels to a tweet exhibited a more coherent and meaningful pattern. A selection of instances of the model's reasoning is provided in Table 2 for reference.

Table 2
Examples of LLM Outcomes

Tweet Text	**Predicted Labels**	**Explanation Generated**
study link hpv vaccine to historically high infertility rate #spirituality #self	['side-effect', 'religious']	*The tweet expresses concern about the potential side effects of the HPV vaccine, specifically linking it to a historically high infertility rate. The tweet also mentions the keywords 'spirituality' and 'self', which suggests that the tweet is expressing opposition to the vaccine based on religious beliefs.*
am apreciat un videoclip pe treatment and cure for gardasil hpv vaccine injury here leylas story	['side-effect']	*The tweet mentions "HPV vaccine injury" and "Leylaś story" which suggests that the tweet is expressing worries about the adverse effects of the HPV vaccine. The keywords "injury" and "story" are associated with the "Side-effect" label, which is why this label is applicable to the tweet.'*
oral polio vaccine infecting unvaccinated kid via #vet #christian #writer #scifi #trump #maga	['ingredients', 'religious']	*The tweet mentions the words 'oral polio vaccine' and 'unvaccinated kid', which suggests that the tweet is raising concerns about the ingredients present in the vaccine and its potential effects on unvaccinated children. The tweet also includes the words 'Christian' and 'Trump', which implies that the tweet is expressing opposition to the vaccine based on religious beliefs. The keywords 'oral polio vaccine', 'unvaccinated kid'*

5.1.1. Template

We used the `text-davinci-003` engine which belongs to the GPT 3.5[3] family, with configuration `temperature` = 0 and `max_tokens` = 100. Our prompt template is outlined as follows:

```
prompt = f"""
    Task: You are given a tweet on vaccines delimited by triple backticks(```). Your task is to pick the right Concern Labels for the tweet from the provided options. Also, explain the logical reasoning you took to choose a label, including the relevant words from the tweets. Learn from examples by understanding common words, keywords, and feelings associated with each label. Only choose labels you're confident about. Each label has a description and common keywords to help you understand the concern that the tweet is depicting.

    Concern List:
    1. Concern 1
    2. Concern 2
    ...

    ###
    Description of the 12 different vaccine concerns :
    1. Concern 1: Description of the Concern.
    ...

    ###
    Some of the keywords associated with tweets for each label:
    1. Concern 1: Keywords - present in a tweet for Concern 1.
    ...

    ###
    Examples:
    Tweet: "Training tweet "
    Concern: ['associated concerns']
    ...

    ###
    Format of response (Response should include Concern in the same format as in examples):
    Concern: [List of all the relevant applicable concern labels]
    Reasoning: [logical reasoning followed to decide each of the applicable labels]

    Tweet: ```{Tweet}```

    Note: Include only the most relevant concern labels in your response. Understand and analyze the sentiment and hidden meanings associated with the given tweet and compare it with the sentiments and keywords in the examples before responding. Comprehend the descriptions and keywords associated with each concern label and then assess the similarity with the given tweet's meaning and these concerns. Verify each and every label before responding to increase the prediction accuracy.
    """
```

5.1.2. Proposed Prompt Template Approach

In the context of addressing a 12-class multi-label classification problem, we propose a structured approach that includes the following essential components:

Task Specification: It is imperative to commence by precisely articulating the classification task under consideration. In our case, this involves the categorization of Social Media Tweets concerning vaccines into multiple concern categories.

Option Presentation: Subsequently, we put forth a comprehensive array of the 12 distinct concern labels for the model's consideration.

Explanatory Context for Options: To facilitate the model's comprehension, we provide detailed explanations and descriptions for each of the concern labels. These descriptions not only convey the essence of each concern but also elucidate common keywords (as per table 6[1] along with some more additions to it), sentiments, or themes that are typically associated with them.

Learning Through Examples: An integral aspect of our approach involves training the model through illustrative examples. These examples consist of actual tweets paired with their corresponding concern labels. Additionally, we furnish comprehensive explanations accompanying these examples, elucidating the logical reasoning behind the assignment of specific labels to each tweet.

In the pursuit of optimizing the model performance, we maintain records of the specific adjustments and enhancements deemed necessary after testing. These modifications include various aspects, such as refining keyword lists, fine-tuning model parameters, and adapting the training strategy.

It is important to emphasize that our methodology transcends mere label assignment, we challenge the model to not only provide labels but also substantiate its choices through logical reasoning or explanations. This approach fosters a deeper understanding of the model's decision-making process, enabling us to elucidate the thought patterns underlying its multi-label classification predictions.

5.2. Transformer based models

5.2.1. DistilBERT & BERT

Distillable Bidirectional Encoder Representations from Transformers is a widely used NLP model. DistilBERT base uncased[4] (which has been used) is a lighter and faster variant of the BERT model that retains much of its capability to understand and generate human-like text.

One-hot encoding was applied on the pre-processed train set. For the first run, the 9921 tweets from the original training dataset were well-shuffled and split into training and validation sets in 80:20 ratio. The 7936 tweets and their ground truth labels were used to fine-tune the pre-trained model and the 1985 tweets (validation data) were used for evaluation purpose. Then, this model was run to predict the classes for the test set, with the hyper-parameters: `max len` = 512, `train batch size` = 16, `learning rate` = 1e-05, `num workers` = 2 and no. of `epochs` = 10.

For the second run, the model was fine-tuned with the entire train set data (without any split). `Dropout rate` of 0.5, `weight decay` of 0.001, `learning rate` of 1e-4, `batch size` 16 and `threshold value` of 0.5 were applied. Model was trained for 10 `epochs`. These hyper parameters values were carefully chosen to collectively contribute to improving the model training efficiency & performance (by mitigating overfitting etc.).

BERT base uncased[5] was also tried (on 80% training & 20% validation data) with the hyper-parameters: `max len` = 200, `train batch size` = 8, `learning rate` = 1e-05 and no. of epochs = 10. BERT base uncased outperformed DistilBERT on the validation set, achieving a macro F1 score of 0.83 after 10 epochs of training. However, when applied to the test set, BERT faced challenges in predicting classes for a significant number of cases, unlike DistilBERT. The superior performance on the validation set can be attributed to BERT's larger capacity to learn from the training data. Yet, this advantage might've also made BERT more susceptible to overfitting, where it closely tailors its predictions to the training data, making it less adaptable to the new test data. It's possible that the hyper-parameters chosen during training favored BERT's performance on the validation set but were less suitable for the test set.

Table 3
Results on Validation Set (20% of training data)

Model	Macro P	Macro R	Macro F1
DistilBERT base uncased	0.80	0.70	0.74

5.3. Traditional Methods

We also explored some of the traditional machine learning techniques like Multinomial Naive Bayes (NB), Random Forest and Support Vector Machine (SVM) with TF-IDF vectorization on this task. However, it became evident during our experiments that these traditional methods struggled to effectively capture the complexity of the problem. The conclusion was drawn based on the validation scores, which consistently demonstrated limitations in their performance.

Table 4
Results on Validation Set (20% of training data)

Model	Macro P	Macro R	Macro F1
SVM with TF-IDF	0.79	0.34	0.45
MN-NB	0.64	0.31	0.39
RF with TF-IDF	0.70	0.18	0.26

5.3.1. Feature extraction

TF-IDF: Term Frequency-Inverse Document Frequency is a widely used text vectorization technique in NLP and information retrieval. It helps in capturing the importance of words in a document (d) and across a corpus. TF measures the frequency of a term (t) within a document while IDF measures the importance of a term across a collection of documents.

$$TF(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d} \tag{1}$$

$$IDF(t,D) = \ln\left(\frac{\text{Total number of documents in corpus } D}{\text{Number of documents containing term } t}\right) \tag{2}$$

TF-IDF is the product of TF & IDF which assigns a weight to each term in a document. Rare words that appear in specific documents will have a high TF-IDF score.

$$TF - IDF(t, d, D) = TF(t, d) \cdot IDF(t, D) \tag{3}$$

5.4. Classifier Chains

Classifier chains[6][7] is a machine learning method for problem transformation in multi-label classification. It is an extension of Binary Relevance, where each label is treated as a separate binary classification problem.

Table 5
Results on Validation Set (20% of training data)

Model	Macro P	Macro R	Macro F1
Classifier Chains with TF-IDF	0.74	0.42	0.48

6. Results

6.1. Metrics

Precision (P): It is the measure of the accuracy of positive predictions made by the model. It can be expressed as

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

where TP and FP refer to True positives (number of correctly predicted positive instances) and False positives (number of incorrectly predicted positive instances) respectively.
Recall(R): It answers what proportion of actual positives was identified correctly. It can be expressed as

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

where FN refers to False negatives (number of incorrectly predicted negative instances).
F-score(F1): It is the harmonic mean of precision and recall. It is a useful metric as it provides a more balanced summarization of model performance, considering both the P and R values. It can be expressed as

$$F1 = 2 \times \frac{(P \times R)}{(P + R)}$$

In terms of TP, FP and FN, the alternative equation can be

$$F1 = \frac{TP}{(TP + \frac{1}{2} \times (FP + FN)}$$

These values are scaled between 0 and 1, with 1 signifying the highest achievable score and 0 denoting the poorest performance.

Jaccard Score: Also known as the Jaccard Index or Jaccard Similarity Coefficient, it is a measure of the similarity between two sets. It is defined as the size of the intersection of the sets divided by the size of the union of the sets. In context of our classification task, it can be represented as:

$$J(y_{\text{true}}, y_{\text{pred}}) = \frac{|y_{\text{true}} \cap y_{\text{pred}}|}{|y_{\text{true}} \cup y_{\text{pred}}|}$$

where `y_true` and `y_pred` represent the ground truth & predicted labels respectively. The Jaccard Score ranges between 0 and 1 where a score of 1 indicates that the predicted labels perfectly match the true labels (there are no false positives or false negatives) & 0 means that there is no overlap between the predicted and true labels, indicating complete dissimilarity.

6.2. Final Outcomes

Out of all the methods we experimented with, the ones sent to the track include the final predictions made by the LLM (run 3) and DistilBERT (runs 1 & 2, as mentioned in 5.2.1) on the test set.

The macro F1[8] and Jaccard[9] scores from the scikit-learn library are the metrics used to evaluate the performance of the model. The scores are computed based on the predicted labels on the test set. LLM achieved a slightly higher macro F1 score of 0.55, outperforming the transformer models like DistilBERT, albeit with a narrow margin.

Coming to the class-wise stats, we see that most of the models tend to struggle with two specific classes, 'none' and 'conspiracy'. Additionally, the 'country' and 'religious' classes, which have relatively fewer examples in the dataset, also tend to result in lower model performance.

We will be able to share the results for our other model predictions on the test data once we have the ground truth labels (which will be revealed after the conference concludes) for the test set.

Table 6
Final results on Test Set

Run File	Method	Macro F1 score	Jaccard score
DataWarriors_Run3.csv	LLM	0.55	0.47
DataWarriors_Run2.csv	DistilBERT (no validation split)	0.54	0.55
DataWarriors_Run1.csv	DistilBERT (with 20% validation split)	0.53	0.56

7. Limitations of the Study & Future Aspects

- Due to constraints of not having an OpenAI paid subscription, the model was trained using a subset of only 58 cases randomly chosen (thoughtfully selected in a way such that examples from all 12 classes were present) from the pre-processed train set. In the future, a more extensive exploration and understanding of diverse prompting strategies could yield even better results. *It's important to highlight that LLM has proved it's effectiveness in terms of performance on this complex multi-label classification task. This suggests that* *with additional resources, there's a significant potential for further enhancing the model's performance.*

- The exploration of advanced models, spanning a broader range of epochs and configurations, was hindered by limitations in GPU resources.
- We observed that the model adhered to the provided concern labels. The structured prompts, incorporating keywords and notes, proved highly effective in reducing hallucinations. This approach empowered the model to develop a nuanced understanding of each concern label, and its responses were logically aligned with this comprehension. However, there are instances where the model exhibited hallucinations while generating explanations - false yet credible-sounding content. The fluency and quality of the generated non-factual content by the models are intriguing & require further study to deduce patterns and understand the model's thought process. The findings from (Augenstein et al., 2023)[10] shed light on similar concerns related to large language model behaviour.
- The incorporation of sentiments into responses, particularly within the domain of large language models, stands as an unexplored area. While it's a preliminary investigation, it provides insights into possible implications and suggests areas for future research.

Acknowledgments

We would like to extend our sincere gratitude to the organizers of the Artificial Intelligence on Social Media (AISoMe) track[11] at the FIRE 2023 conference, for hosting this brilliant & insightful problem statement that served as a great learning experience. We're grateful for the opportunity to engage with it.

References

[1] S. Poddar, A. M. Samad, R. Mukherjee, N. Ganguly, S. Ghosh, CAVES: A Dataset to facilitate Explainable Classification and Summarization of Concerns towards COVID Vaccines, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 3154–3164.
[2] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, A. Kalyan, Learn to explain: Multimodal Reasoning via Thought Chains for Science Question Answering, in: The 36th Conference on Neural Information Processing Systems (NeurIPS), 2022.
[3] Website. URL: https://platform.openai.com/docs/models/gpt-3-5.
[4] Website. URL: https://huggingface.co/distilbert-base-uncased.
[5] Website. URL: https://huggingface.co/bert-base-uncased.
[6] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier Chains for Multi-label Classification, 2009. URL: https://www.cs.waikato.ac.nz/~eibe/pubs/chains.pdf.
[7] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains: A Review and Perspectives, 2020. URL: https://doi.org/10.48550/arXiv.1912.13405.
[8] Website. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html.

[9] Website. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.jaccard_score.html.

[10] I. Augenstein, T. Baldwin, M. Cha, T. Chakraborty, G. L. Ciampaglia, D. Corney, R. DiResta, E. Ferrara, S. Hale, A. Halevy, E. Hovy, H. Ji, F. Menczer, R. Miguez, P. Nakov, D. Scheufele, S. Sharma, G. Zagni, Factuality Challenges in the Era of Large Language Models, 2023. arXiv:2310.05189.

[11] S. Poddar, M. Basu, K. Ghosh, S. Ghosh, Overview of the FIRE 2023 Track: Artificial Intelligence on Social Media (AISoMe), in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, 2023.

[12] B. Zhang, D. Ding, L. Jing, How would Stance Detection Techniques Evolve after the Launch of ChatGPT?, 2023. arXiv:2212.14548.