Multi-Dimensional Classification on Social Media Data for Detailed Reporting with Large Language Models

Riccardo Cantini[1][0000−0003−3053−6132], Cristian Cosentino[1][0000−0002−6368−373X], and Fabrizio Marozzo[1][0000−0001−7887−1314]

University of Calabria, Italy
{rcantini, ccosentino, fmarozzo}@dimes.unical.it

Abstract. Every day, more and more people harness the power of social media platforms to express their thoughts, share information and personal experiences, and engage with others. All this knowledge can then be transformed into informative reports with the assistance of Large Language Models (LLMs), like ChatGPT, which leverage deep learning techniques to analyze data and generate comprehensive analyses. By effectively classifying user-generated posts based on dimensions such as topic, sentiment, and emotion, it is possible to create even more detailed reports by carefully condensing large amounts of data collected along the different dimensions considered. To tackle this challenge, we have developed an automated approach with two primary goals: (i) categorizing posts across different dimensions using ready-to-use and fine-tuned classifiers; and (ii) generating detailed reports via LLMs that summarize posts with similar characteristics along the defined dimensions. In our analysis, we examined a large and varied set of posts about COVID, classifying them along several dimensions, including topic, content type, expressed sentiment and emotions, and reliability of information. Specifically, by choosing to generate a report for the main discussion topics present in the dataset, such as allergic reactions or school issues, and using the remaining dimensions for post classification, we successfully created highly detailed and informative reports with ChatGPT. These reports outperformed those generated directly by ChatGPT, in both quantitative measures such as linguistic scores and qualitative evaluations by field experts.

Keywords: Large Language Models · Deep Learning · Natural Language Processing · ChatGPT · Social media data · Reporting.

1 Introduction

In today's digital age, social media has become an essential component of our lives, radically transforming communication, information sharing, and global interactions [3]. With the continuously growing number of active users across popular platforms such as Facebook, Instagram, WhatsApp and Twitter, social

media has become an expansive and remarkable repository of data, encapsulating personal opinions, diverse perspectives, real-time updates, and global interactions [5]. Such data can be leveraged by both organizations and individuals, allowing them to make informed, data-driven decisions and gain a comprehensive understanding of society at large.

Traditionally, extracting actionable insights from extensive datasets has been a labor-intensive process [4], requiring manual analysis and interpretation of statistical data to generate meaningful reports. However, recent advancements in artificial intelligence (AI) and Natural Language Processing (NLP), particularly Large Language Models (LLMs), have revolutionized data analysis and reporting practices [15]. LLMs excel in understanding and generating human-like text, leveraging deep learning techniques to analyze vast datasets efficiently. This innovation allows for significant time savings, greater precision, and simplified decision-making processes like never before.

In this paper, we introduce a novel methodology to generate high-quality and comprehensive reports from user-generated content on social media platforms. By systematically categorizing posts across dimensions such as topic, content type, sentiment, emotional expression, and the presence of false information, we unlock the potential for in-depth analysis and understanding of aggregated data. By leveraging Large Language Models (LLMs), we further enhance this analysis, enabling the description of results in a human-readable manner.

To achieve this goal, our automated approach comprises two primary objectives. Firstly, we employ ready-to-use and fine-tuned classifiers to categorize posts across various dimensions, including topic, emotion, sentiment, presence of false information, and content type. Secondly, we use LLMs to generate detailed reports summarizing posts with similar characteristics along these defined dimensions. This approach enables us to create comprehensive reports that allow us to deeply understand the patterns, trends, and insights embedded within the user-generated content on social media platforms.

To assess the effectiveness of our methodology, we conducted an extensive evaluation on a large and varied dataset consisting of 15 million COVID-related posts in order to measure the accuracy of categorization and the quality of the reports generated. These posts were classified according to several dimensions, including topic, type of content, expressed sentiments and emotions, and presence of false information. For each topic identified (19 topics extracted), we generated an informative and detailed report. In our evaluation, we found that these reports outperformed those generated directly by Large Language Models (LLMs) without prior categorization along dimensions both with quantitative measures, such as linguistic scores, and qualitative evaluations by field experts.

The paper is structured as follows: Section 2 reviews related work in the fields of report generation with LLMs. Section 3 describes our methodology, while Section 4 presents the achieved results, including sample reports and their evaluation. Finally, Section 5 concludes the paper, summarizing our contributions and discussing avenues for future research.

2 Related Work

With the advent of AI-powered by the latest and most advanced LLMs, the process of extracting insights and knowledge from data has become more streamlined and user-friendly. LLMs are equipped with natural language processing (NLP) and machine learning capabilities, enabling them to understand user queries in plain language and generate textual reports that offer relevant information and insights from the data. These LLMs serve as interactive guides, leading users through the data analysis process and presenting the results in an easily understandable and interpretable manner. They can address specific questions, provide explanations, and even offer visualizations as needed, thereby enhancing the overall user experience.

In current research, LLMs are exploited as versatile tools across several domains, spanning from education to e-commerce, healthcare, and entertainment. LLMs in these domains primarily aid users in information retrieval activities, leveraging their chatbot capabilities [7]. For instance, educational LLMs often provide specific information such as class schedules or educational materials [21]. Similarly, healthcare LLMs predominantly focus on information retrieval for healthcare assistance, utilizing machine learning techniques to assist in disease pre-diagnosis [2,1]. Likewise, e-commerce LLMs are commonly employed to offer customer support, provide information on product catalogs, and enhance the overall shopping experience [18]. Furthermore, even when it comes to crucial data annotation tasks in numerous strategic and industrial sectors, LLMs have shown superior performance compared to crowd-workers in areas such as relevance, positioning, topics and frame detection [13].

In the realm of report generation, text models like GPT demonstrate versatile applications, expanding into novel contexts and showcasing their adaptability. For instance, Wang et al. [22] address the challenge of controlled text generation from tables, aiming to provide natural language descriptions of specific sections highlighted within tables. Their approach prioritizes content relationships over sequential generation, enhancing model robustness against structural variations while ensuring accurate descriptions of highlighted table sections. Messina et al. [20] conduct an in-depth analysis of deep learning-based methods for automatically generating reports from medical images. They illustrate the integration of deep learning algorithms for image analysis with natural language processing techniques for report writing, exemplified by concise X-ray reports. Belcastro et al. [6] propose a methodology for interpretable depression detection, combining Large Language Models (LLMs) with eXplainable Artificial Intelligence (XAI) and conversational agents like ChatGPT. This approach facilitates the generation of reports regarding mental disorders expressed in social media posts. In the finance domain, GPT models are utilized for generating financial reports, summaries, and forecasts, as well as text-based financial analysis including sentiment, news, and social media analysis. Zaremba et al. [23] highlight their accuracy and reliability in providing financial information. Dwivedi [12] underscores the benefits of ChatGPT adoption by regulators in the tourism and hospitality industry, enabling efficient extraction of information about organizations, compliance in-

vestigations, and real-time explanations. Moreover, in Information Technology, particularly in log analysis, ChatGPT proves useful for summarizing logs and improving their organization and comprehensibility due to its expository capability. Meng et al. [19] introduce LogSumm, which simplifies the synthesis of crucial information contained in logs, thereby enhancing data organization and comprehension through automation

In contrast to previous works, our study focuses on assessing how categorizing input contents can significantly enhance the quality of reports generated by Large Language Models (LLMs) from textual data, such as those sourced from social media users. Furthermore, we show methods to evaluate the quality of these generated reports and conduct comparative analyses between them.

3 Proposed methodology

The proposed methodology is aimed at transforming vast amounts of user-generated posts into informative reports by effectively categorizing them according to various dimensions like topic, sentiment, and emotion. It comprises three distinct phases: (i) *collecting posts and classifying them along dimensions*; (ii) *grouping similar posts on dimensions*; and (iii) *summarizing and reporting through LLMs*. In the following, we provide a detailed description of the main steps of our approach, whose execution flow is depicted in Figure 1

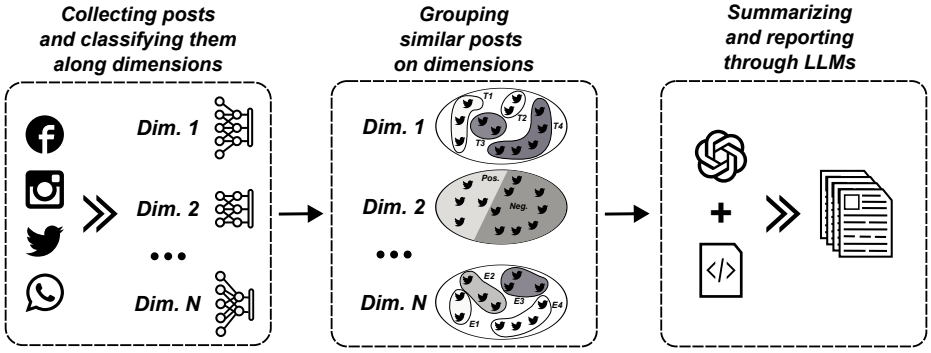


Fig. 1. Execution flow of the proposed methodology.

In the initial phase of our methodology - *collecting posts and classifying them along dimensions* - we undertake a thorough process to gather user-generated content from various social media platforms. Each post undergoes in-depth classification across multiple dimensions, such as topic, information type, sentiment, emotion, and false information. To achieve accurate categorization, we select the most effective classifiers available in the literature for each dimension. This careful selection ensures the accuracy of our classification process, which is crucial for the subsequent phase of grouping similar posts together. Our classifiers include

a wide range of techniques, extending from standard models to sophisticated pre-trained models such as BERT-based architectures [11]. Furthermore, these models offer flexibility, as they can be utilized as-is or undergo additional training to adapt them to specific tasks and/or data. For example, in the task of identifying topics [9], BERTopic can be directly applied to textual data, thanks to its efficacy in capturing topic structures without requiring specialization or adaptation to a specific data. Conversely, achieving optimal performance in discerning the type of information (e.g., news or opinion) or distinguishing between reliable and misleading data with BERT-based models necessitates fine-tuning on specific datasets.

In the second phase of our methodology, which is centered on *grouping similar posts on dimensions*, we systematically aggregate posts sharing common features within each defined dimension. For instance, in the sentiment dimension with its two classes (positive and negative), we gather all positive posts into one grouping, and negative ones into another one. The same approach is applied to other dimensions, each of which may have two or more classes. It's important to note that, rather than simply assigning posts to classes, these algorithms often provide probabilities indicating the likelihood of belonging to various classes. Additionally, in our preparation of inputs for Large Language Models (LLMs), we prioritize the inclusion of essential information while considering the limited capacity of input prompts in terms of tokens (e.g., approximately 4096 tokens for ChatGPT-3.5 models). By leveraging this probabilistic information, we strategically select the most representative posts for each class (i.e., those with higher scores), ensuring a balanced and informative representation within the groupings. This selective approach ensures that we provide LLMs with relevant details without overwhelming them, thereby optimizing the effectiveness of our analyses.

In the final phase of our methodology, focused on *summarizing and reporting through LLMs*, we leverage the capabilities of Large Language Models (LLMs) to extract actionable insights from the grouped posts. By using the available APIs, we transmit data grouped along different dimensions to the LLMs and request summaries for each group of similar posts. This method ensures that the LLMs concentrate on relevant subsets of data, optimizing their efficiency and effectiveness in generating summaries. These summaries encapsulate the main themes, sentiments, and insights derived from each group, offering concise and informative snapshots of the user-generated content on a given dimension. This approach not only streamlines the summarization process but also enables a more tailored and targeted analysis. By aggregating summaries from different dimensions, we compile a comprehensive final report that provides a broad and detailed perspective on various aspects of interest. In doing so, we furnish the information necessary for informed decision-making, ultimately transforming individual, timely user-generated content into valuable, actionable aggregate information.

4 Experimental Results

The case study we considered relates to tweets generated during the COVID-19 pandemic. These tweets are used in several research works due to their inclusion of opinions, discussions, the presence of false information, and more. Specifically, we analyzed a vast dataset of 15 million COVID-related posts [17], employing classifiers tailored to different dimensions such as topic, sentiment, and emotion. Given the truly large and varied nature of the dataset, we focused on a subset of the data generated in January 2021, encompassing 303,541 tweets. Our objective is to generate a series of reports derived from users' tweets, describing discussion topics, informative content, expressed opinions, emotions, and the presence of false information.

In the following sections, we comprehensively discuss the experimental results we achieved, with a focus on: (*i*) the selection of dimensions and the identification of the most suitable models for classifying posts according to those dimensions; (*ii*) the grouping of similar posts and the formulation of prompts to summarize their main characteristics; and (*iii*) the generation of reports and a comparison of our methodology with a basic approach.

4.1 Dimension definition and post classification

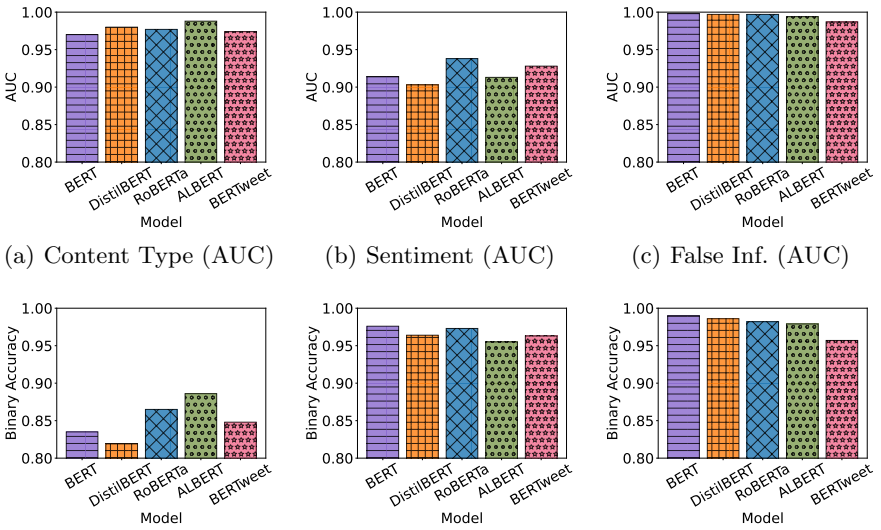
The selection of dimensions heavily relies on the nature of the data under analysis. Given our focus on COVID-19, we identified the following five dimensions (with relative classes) that are pertinent to our case study:

1. *Topic*: indicates the subject matter discussed in a post (classes cannot be defined apriori).
2. *Content Type*: distinguishes between informational content, such as news articles, and personal opinions expressed by users (classes: *news* or *opinion*).
3. *Sentiment*: determines whether a post conveys a positive or negative sentiment (classes: *positive*, *negative*).
4. *Emotion*: identifies the emotional tone and expressions conveyed within the text (classes: *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, *trust*).
5. *False Information*: classifies posts as either disseminating false information or providing reliable content (classes: *reliable information*, *false information*).

To ensure accurate categorization, we chose the most accurate classifiers from the existing literature for each dimension. For *topic* extraction, our selection was BERTopic, as recommended in [14], where it demonstrates better performance compared to other techniques in terms of both topic coherence and diversity. For *emotion*, we used the NRCLex tool [1], a comprehensive reference library capable of identifying a broad spectrum of emotions conveyed in the text.

While for *topic* and *emotion* we used models as-is, for other dimensions we needed models with additional training to adapt them to specific tasks and/or

[1] `https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`



(a) Content Type (AUC) (b) Sentiment (AUC) (c) False Inf. (AUC)

(d) Cont. type (Bin. Acc.) (e) Sentiment (Bin. Acc.) (f) False Inf. (Bin. Acc.).

Fig. 2. BERT-based model comparison in terms of AUC and binary accuracy.

data. In particular, to address the remaining dimensions, we used BERT-based models [11], due to their effectiveness in capturing semantic and syntactic features of microblog texts [10]. Specifically, for the *content type* dimension, we fine-tuned BERT-based models on a dataset comprising 30,000 tweets classified as either informational content (e.g., news articles) or personal opinions [2]. For the *sentiment* dimension, we utilize a training dataset of 16,000 tweets classified as positive or negative [3]. Regarding the *false information* dimension, we leverage a dataset consisting of 15,000 tweets categorized as either false information or reliable content specific to COVID-19 [8].

Figure 2 visually presents the performance analysis of a range of BERT-based models, namely BERT, DistilBERT, RoBERTa, ALBERT, and BERTweet, evaluated in terms of both Area Under the Curve (AUC) and binary accuracy metrics. The observed differences in performance, while small, are attributed to the nuanced architectures and training methodologies of each model. In our analysis of content classification, ALBERT emerges as the most proficient model, albeit marginally, showcasing superior capability. On the other hand, when it comes to sentiment and misinformation classification tasks, RoBERTa and BERT exhibit notable effectiveness, outperforming other counterparts in these specific domains.

Table 1 presents five tweet examples classified across different dimensions. The first example illustrates false information regarding allergic reactions to

[2] https://www.kaggle.com/datasets/ferno2/training1600000processednoemoticoncsv
[3] https://www.kaggle.com/datasets/vinayakshanawad/us-news-dataset

vaccines, expressed with negative sentiment and anger. The second example expresses disappointment regarding vaccine accessibility for individuals with allergies, categorized under the topic of allergic reactions. The third and fourth examples discuss the safety of COVID-19 vaccines during pregnancy and breastfeeding, conveying positive sentiments with anticipation and trust, respectively. Lastly, the fifth example is a positive tweet emphasizing trust in continued preventive measures despite vaccination, falling under the topic of masks. These examples provide insights into the diverse discussions surrounding COVID-19 vaccines across various topics, sentiments, and information reliability.

Table 1. Example of tweets classification on different dimensions (topic, content type, sentiment, emotion and false information).

ID	Tweet	Topic	Content t.	Sentiment	Emotion	False inf.
1	*The vaccine leads to serious allergic crises maybe it's better not to get vaccinated you risk your life less*	Aller. react. (0.98)	Opinion (0.80)	Negative (0.60)	Anger (0.31)	False (0.98)
2	*Disappointed to hear my nan and grandad weren't given their pfizer covid vaccine as they are both allergic to penicillin... better safe than sorry*	Aller. react. (0.98)	Opinion (0.95)	Negative (0.65)	Sadness (0.25)	Reliable (0.90)
3	*No data to suggest mrna vaccines affect breast milk significantly. #coronavirus vaccine may be offered to lactating persons at high risk for infection*	Pregancy and breastfeeding (0.92)	News (0.99)	Positive (0.55)	Antic. (0.28)	Reliable (0.76)
4	*this is The right path. the fda approved a covid19 vaccine and in their approval they also open for pregnant and breastfeeding women to receive it*	Pregancy and breastfeeding (0.97)	News (0.91)	Positive (0.62)	Trust (0.20)	Reliable (0.78)
5	*Remember the vaccine is not immunity, wear a mask*	Mask (0.93)	Opinion (0.99)	Positive (0.62)	Trust (0.31)	Reliable (0.90)

4.2 Post grouping and report generation

In this section, we describe how posts are grouped about our case study and how reports are generated. Posts can be grouped based on their content and the classification assigned to each dimension. The groups for the defined dimensions were previously outlined in the preceding section. The only dimension for which we were unable to define a priori classes is the topic dimension. The classes extracted using BERTopic are 19 in total, annotated according to the specifications in paper [8]. These classes are: allergic reactions, microchip vaccine, lockdown, pregnancy and breastfeeding, Covid European Union, B.Johnson (vaccination), masks, Johnson & Johnson, Pfizer vaccine, Dr. A.Fauci, J.Biden (vaccination), Covid vaccines, 2021 new year, Covid old people, passports vaccine, school issues, covid Florida, workers and employers, olympic games and NBA.

We established a structured report definition by assigning each extracted topic its own report. As our Language Model (LLM), we chose ChatGPT, specifically the ChatGPT 3.5 version. For each topic, we collect all the tweets associated with it, and after classifying these tweets on different dimensions, we

interact with a sequence of prompts with ChatGPT to generate various outputs: *i*) `title_prompt` to generate the title of the report; *ii*) `introduction_prompt` to create the introduction paragraph using informative tweets; *iii*) `sentiment_prompt` to generate a paragraph describing the positive and negative opinions expressed by users; *iv*) `emotion_prompt` to generate a paragraph describing the emotions expressed by users; and *v*) `falseinformation_prompt` to analyze which arguments are mainly influenced by false information.

For each request, we provided a collection of representative tweets for each dimension or class. For example, to address the dimension of sentiment and its positive class, we selected the tweets with the highest positive scores, determining their number based on the token limit of the prompt, which defines its capacity in terms of words. ChatGPT models typically have a maximum token limit of around 4096 tokens for GPT-3.5 models. The task assigned to ChatGPT is as follows:

`behavior`: Act as an expert writer creating reports based on a series of tweets.

The sequence of prompts are defined as follows:

`title_prompt` = *Generate a concise and captivating English title for the report, within a maximum of 10 words. Avoid the use of words with hashtags(#). Use the following tweets as input:* {*news_tweets*}.

`introduction_prompt` = *Generate a concise introduction for the report. It must be interesting and engaging, capture the reader's attention and present key information. It must be a single paragraph, without carriage returns and the use of colons (:). Do not report the content of tweets. Use the following tweets as input:* {*news_tweets*}.

`sentiment_prompt` = *Generate a section about positive and negative tweets written by users. It must be a single paragraph [...]. Use* {*positive_tweets*} *as positive tweets and* {*negative_tweets*} *as negative tweets.*

`emotion_prompt`: *Generate a section that analyzes the emotion expressed in the tweets written by users. It must be a single paragraph [...]. Use the following emotion/tweets pairs as input:* {*anger: anger_tweets*}, {*anticipation: anticipation_tweets*}, *[...]* {*trust: trust_tweets*}.

`falseinformation_prompt`: *Generate a section to analyze false information in tweets. It must be a single paragraph [...]. Use the following false information tweets as input:* {*falseinformation_tweets*}.

These prompts generate five textual outputs (e.g., `title_output` for `title_prompt`) by using the tweets appropriately grouped. After generating the different outputs, we aggregate them together with the following prompt:

`report_prompt`: Generate the final report. Use '`title_output`' as title . Write four distinct paragraphs: the first paragraph by using '`introduction_output`', the second paragraph with '`sentiment_output`', the third paragraph with '`emotion_output`', and the fourth paragraph with '`falseinformation_output`'. Do not separate the

text into sections but write a report composed of 4 paragraphs. Do not use bulleted and numbered lists, and colons(:). Avoid text replies and keep the report within {N} words."

4.3 Performance comparison

In this section, we describe the evaluation of our methodology's performance compared to a standard approach. Specifically, we compare the effectiveness of standard ChatGPT in generating reports versus ChatGPT enhanced with Multi-Dimensional Categorization (*ChatGPT-MDC*).

When using standard ChatGPT for report generation, we adhere to the same procedure outlined in the previous section. However, the prompts provided to standard ChatGPT utilize uncategorized and generic tweets. For instance, when requesting an analysis of positive and negative posts, we provide generic tweets and instruct ChatGPT to identify any positive or negative aspects present within them.

For each test, we generated two reports of approximately 1000 words (around two pages) using standard ChatGPT and ChatGPT-MDC. Below, we offer a comparison of two reports generated on the topic of *allergic reactions*. Due to space limitations, we are unable to present the full reports. However, our objective is to highlight the key aspects of each report to demonstrate the differences in approach between the two methods.

`ChatGPT_output` = "Allergy Warnings Surrounding Pfizer's COVID-19 Vaccine"

The Pfizer COVID-19 vaccine has been making headlines worldwide, with reports of allergy warnings and adverse reactions surfacing. [...]

Positive and negative sentiments surrounding the Pfizer COVID-19 vaccine have been expressed on social media platforms. Users have shared their experiences and concerns regarding allergic reactions to the vaccine. Some users have reported their hospitalization soon after receiving the vaccine. Others question the lack of reported allergic reactions beyond the initial incidents, suggesting a need for more information. [...]

The sentiment analysis of the tweets surrounding the allergy warnings of the Pfizer COVID-19 vaccine reveals a mixture of concerns, frustrations, and questions from the public. Some individuals express frustration and disappointment, feeling that their allergies may prevent them from returning to a normal life. [...]

False information about allergy warnings for the Pfizer COVID-19 vaccine is circulating on social media, containing misleading statements. ... Some tweets suggest allergic reactions are only in those with previous allergies, but reactions vary individually. [...]

`ChatGPT-MDC_output` = "Allergic Reactions to Pfizer's COVID-19 Vaccine: Risk and Warnings"

The Pfizer-BioNTech COVID-19 vaccine has been widely administered across the globe as part of the ongoing efforts to combat the pandemic. [...]

Positive tweets expressed optimism and confidence in the vaccine's safety and efficacy. Users highlighted the speed of vaccine distribution, with one tweet mentioning the clearance of the vaccine within 24 hours. Another tweet shared an article discussing what people with allergies should know about COVID vaccines, indicating a desire for accurate information. [...] However, a significant number of negative tweets expressed worry and skepticism about the vaccine's potential side effects, particularly for individuals with allergies. [...]

The sentiment analysis of tweets regarding allergic reactions to Pfizer's COVID-19 vaccine reveals a range of emotions among the public. There is a notable sense of fear and concern expressed in tweets discussing the allergy warning and the investigation into allergic reactions. Some individuals express anger, questioning the blame placed on allergies for healthcare workers' symptoms after vaccination. Anticipation is also evident, with tweets acknowledging potential allergic reactions but urging the public not to be overly anxious. [...]

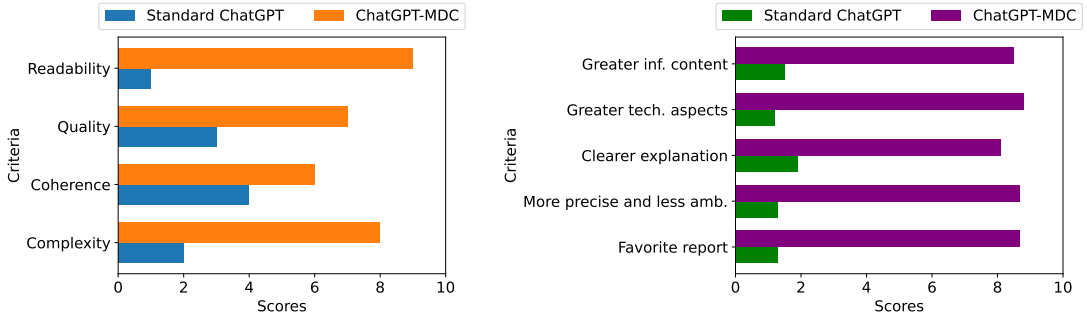
Regarding the dissemination of false information, our analysis found that 36.8% of the examined tweets contained misleading claims. These tweets claimed that the vaccines were coated with a toxic chemical that could prove deadly, citing exclusive reports. It is important to note that these claims are unsubstantiated and lack scientific evidence. [...]

Observing the comparison, the standard ChatGPT approach offers vague descriptions of the various sections of the report, lacking specific information. In contrast, the enhanced approach of ChatGPT-MDC, equipped with categorized tweets for each dimension and associated classes, demonstrates notably greater precision. Not only does it provide concrete examples, but it also furnishes statistical data and comprehensive analyses, particularly evident in the examination of various expressed emotions.

However, to measure the quality of the reports generated by the two methods, we used two approaches:

1. The TextDescriptives library [16] was employed to analyze reports and generate statistics from text. Specifically, diverse scores were utilized to assess the readability, quality, coherence, and complexity of the generated reports.
2. Twenty domain experts were enlisted to assess the generated reports. Their task involved comparing two reports on the same topic, responding to specific questions, and determining which report excelled in terms of informative content, technical information, textual clarity, and precision.

Figure 3(a) reports scores derived from generated text on ten topics using TextDescriptives. Based on the provided scores obtained from the TextDescriptives library on reports generated by ChatGPT and ChatGPT-MDC, here is a description of each criterion and the results obtained:



(a) TextDescriptives. (b) Domain expert.

Fig. 3. Scores obtained in the evaluation of reports with the TextDescriptives library and by domain experts.

- The *Readability* of the reports was assessed using the Coleman-Liau index, which estimates the U.S. grade level required to understand a text. Reports generated by ChatGPT-MDC require a higher U.S. grade level (around 18) compared to those generated by ChatGPT (around 16). Other readability indices such as Gunning-Fog and SMOG also indicate similar trends, suggesting that the reports from ChatGPT-MDC required a deeper understanding of linguistics.
- *Quality* was measured using repetitive text metrics, specifically the duplicate n-gram character fraction, which indicates the fraction of characters in a document that are contained within duplicate n-grams. Reports generated by ChatGPT-MDC exhibit less repetitiveness, implying more informative content compared to reports from ChatGPT.
- *Coherence* was evaluated based on the cosine similarity between sentences, whose embedding was obtained as the average vector representation of words computed by Latent Semantic Analysis. Reports from ChatGPT-MDC generally demonstrate higher coherence, as indicated by higher first-order coherence values (cosine similarity between consecutive sentences).
- *Complexity* was assessed using the entropy of the text, which measures the level of randomness or unpredictability, with higher values indicating greater diversity and complexity of language use. Reports from ChatGPT-MDC demonstrate higher complexity, characterized by greater diversity and complexity of language use, compared to those from ChatGPT, which exhibit more repetitive or predictable language patterns.

Regarding the evaluations by domain experts, we conducted a test consisting of ten questions. In each test, we presented complete reports (or excerpts) and asked them to identify which report excelled in specific aspects. Specifically, they were asked to answer the following questions: (*i*) which report do you believe offers greater overall information content? (*ii*) which report contains more technical or specialized aspects? (*iii*) which report provides a clearer explanation of

the topics covered? (*iv*) which report demonstrates greater precision and clarity in its contents? (*v*) which report do you prefer for overall quality?

Figure 3(b) illustrates the results obtained on a scale from 0 to 10 for different evaluation criteria. Domain experts consistently favored ChatGPT-MDC over standard ChatGPT across all aspects. ChatGPT-MDC received significantly higher ratings attributed to its precision, aided by categorized tweets and statistical data. Reports generated by ChatGPT-MDC were notably less vague, more detailed, and effectively utilized numerical data, enhancing clarity and credibility. Furthermore, they exhibited superior coherence, organization, and grammatical consistency, resulting in higher evaluation scores compared to standard report generation with ChatGPT.

5 Conclusion

The findings of our study demonstrate the potential of leveraging Large Language Models (LLMs) such as ChatGPT, in conjunction with Multi-Dimensional Categorization (MDC), to generate highly informative reports from user-generated posts on social media platforms. By effectively categorizing posts along dimensions like topic, sentiment, and emotion, we were able to condense vast amounts of data into detailed reports that provide insights into various aspects of online discourse, such as COVID-related discussions. We plan to broaden our approach to diverse datasets like product reviews and scientific literature, refining categorization methods and utilizing more powerful Large Language Models. This expansion will enhance decision-making, aid research, and extract insights from large-scale textual data.

Acknowledgements

This work was supported by the research project "INSIDER: INtelligent ServIce Deployment for advanced cloud-Edge integRation" granted by the Italian Ministry of University and Research (MUR) within the PRIN 2022 program and European Union - Next Generation EU (grant n. 2022WWSCRR, CUP H53D23003670006). It was also supported by the "National Centre for HPC, Big Data and Quantum Computing", CN00000013 - CUP H23C22000360005, and by the "FAIR – Future Artificial Intelligence Research" project - CUP H23C22000860006.

References

1. Athota, L., Shukla, V.K., Pandey, N., Rana, A.: Chatbot for healthcare system using artificial intelligence. In: 2020 8th Int. Conf. on Reliability, Infocom Technologies and Optimization (Trends and Future Directions). pp. 619–622 (2020)
2. Ayanouz, S., Abdelhakim, B.A., Benhmed, M.: A smart chatbot architecture based nlp and machine learning for health care assistance. In: Proceedings of the 3rd International Conference on Networking, Information Systems & Security (2020)

3. Baym, N.K.: Personal connections in the digital age. John Wiley & Sons (2015)
4. Belcastro, L., Cantini, R., Marozzo, F., Orsino, A., Talia, D., Trunfio, P.: Programming big data analysis: Principles and solutions. Journal of Big Data **9**(4) (2022)
5. Belcastro, L., Cantini, R., Marozzo, F., Talia, D., Trunfio, P.: Learning political polarization on social media using neural networks. IEEE Access **8** (2020)
6. Belcastro, L., Cantini, R., Marozzo, F., Talia, D., Trunfio, P.: Detecting mental disorder on social media: a chatgpt-augmented explainable approach. arXiv:2401.17477 (2024)
7. Caldarini, G., Jaf, S., McGarry, K.: A literature survey of recent advances in chatbots. Information **13**(1) (2022)
8. Cantini, R., Cosentino, C., Kilanioti, I., Marozzo, F., Talia, D.: Unmasking covid-19 false information on twitter: a topic-based approach with bert. In: 26th International Conference on Discovery Science (DS2023). vol. 14276, pp. 126–140 (2023)
9. Cantini, R., Marozzo, F.: Topic detection and tracking in social media platforms. In: EAI PerSoM 2022. pp. 41–56 (2022)
10. Cantini, R., Marozzo, F., Bruno, G., Trunfio, P.: Learning sentence-to-hashtags semantic mapping for hashtag recommendation on microblogs. ACM Transactions on Knowledge Discovery from Data **16**(2), 1–26 (2022)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
12. Dwivedi, Y., Pandey, N., Currie, W., Micu, A.: Leveraging chatgpt and other generative artificial intelligence (ai)-based applications in the hospitality and tourism industry: practices, challenges and research agenda. International Journal of Contemporary Hospitality Management **36** (06 2023)
13. Gilardi, F., Alizadeh, M., Kubli, M.: Chatgpt outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences **120**(30) (2023)
14. Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure (2022)
15. Hadi, M.U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M.B., Akhtar, N., Wu, J., Mirjalili, S., et al.: A survey on large language models: Applications, challenges, limitations, and practical usage. Authorea Preprints (2023)
16. Hansen, L., Olsen, L.R., Enevoldsen, K.: Textdescriptives: A python package for calculating a large variety of metrics from text. Journal of Open Source Software **8**(84), 5153 (Apr 2023)
17. Hayawi, K., et al.: Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. Public health **203**, 23–30 (2022)
18. Hossain, M., Habib, M., Hassan, M., Soroni, F., Khan, M.M.: Research and development of an e-commerce with sales chatbot. In: 2022 IEEE World AI IoT Congress (AIIoT). pp. 557–564 (2022)
19. Meng, W., Zaiter, F., Zhang, Y., Liu, Y., Zhang, S., Tao, S., Zhu, Y., Han, T., Zhao, Y., Wang, E., et al.: Logsummary: Unstructured log summarization for software systems. IEEE Transactions on Network and Service Management (2023)
20. Messina, P., et al.: A survey on deep learning and explainability for automatic report generation from medical images. ACM Comp. Surveys **54**(10s), 1–40 (2022)
21. Okonkwo, C.W., Ade-Ibijola, A.: Chatbots applications in education: A systematic review. Computers and Education: Artificial Intelligence **2**, 100033 (2021)
22. Wang, F., Xu, Z., Szekely, P., Chen, M.: Robust (controlled) table-to-text generation with structure-aware equivariance learning. arXiv:2205.03972 (2022)
23. Zaremba, A., Demir, E.: Chatgpt: Unlocking the future of nlp in finance. Modern Finance **1**(1), 93–98 (2023)