Public Perception of Dark Matter: A Social Media Analysis

Cassandra Chang, Vivian Ha, Diaab Khan, Jenevieve Liem, Phuong Pham, Hugh Qiu

Group 10
Supervisor: Dr Ciaran O'Hare
School of Physics, The University of Sydney, Sydney NSW 2006 Australia

Abstract: Dark matter, comprising about 85 percent of the universe's total matter matter, is a form of invisible mass that surrounds the universe [1]. This study employs AI to analyse discussions on dark matter across social media platforms (Youtube Shorts, Youtube Videos, Tiktok, Reddit, Facebook, Instagram Reels and Instagram Posts) in terms of scientific relevance, polarity, objectivity and accuracy. By examining 140 sources and 7411 comments, our study aims to provide insights that will enhance the scientific communicationof dark matter and also serve as a foundation for future study in science communication. From our extensive analysis, we concluded that Reddit generated the most fruitful scientific discussion. Recommendations include leveraging Reddit for meaningful discussions and providing prompts in posts on other social media platforms to guide discussions in the comments sections. In short-form videos, it is recommended to emphasise consensus and evidence for dark matter to reduce scepticism and negativity. Clear explanations of the specific percentages used to describe dark matter, the fringe status of modified Newtonian dynamics (MOND) in the scientific community and the distinctions between dark matter, dark energy and black holes are also crucial in enhancing the accuracy with which the public discusses dark matter.

1 Introduction

Dark matter is a form of matter only detected through its gravitational influence on the celestial bodies it surrounds [1]. Estimated to constitute approximately 85% of the total matter in the universe, it possesses a unique quality of being unable to reflect, emit or absorb any light or other forms of electromagnetic radiation [1], for a review see Ref [2]. This has ultimately made the process of determining its composition difficult for researchers.

There have been many observations in the past with findings that support the presence of dark matter - some of the most notable being from the investigations of Fritz Zwicky in 1933 as well as Vera Rubin and Kent Ford in the 1970s. They were able to respectively uncover that galaxies' rotational velocities were moving too fast to remain gravitationally bound to their clusters [3] and that outer stars had the same orbital velocity as those closer to the centre [4]. This could not be entirely explained by visible matter, so it was suggested that there are large amounts of "dark mass" in surrounding regions [3][4].

In more recent times, there are two popular candidates to explain the composition of dark matter which include Massive Compact Halo Objects (MACHOS), and weakly interacting massive particles (WIMPs). MACHOs describe compact objects that are similar to ordinary stars but are less luminous while WIMPs are hypothesised to be a class of particles that also do not absorb or emit light, and do not interact strongly with other particles. [5]. There are many experiments dedicated to detecting dark matter including XENONnT, which has been designed to detect WIMPs through nuclear recoil interactions with liquid and gaseous xenon [6].

2 Objective, Significance and Hypotheses

The perceived lack of relevance of Dark Matter to daily life and its current "invisibility" could greatly impact its public perception. This report aims to investigate how the public considers the topic of "Dark Matter" through an AI-powered analysis of social media discussions. In doing so, it will offer insights that will enhance the scientific communication of dark matter and provide a foundation for future investigations regarding the public's comprehension of scientific phenomena. Improving the way we communicate dark matter will encourage new scientists to pursue this field and enhance our understanding of the universe.

This research will assess specific social media platforms in terms of the following parameters: scientific relevance, objectivity, polarity (positivity) and accuracy. As a result, this study will give a sophisticated and detailed picture of the discussion related to ”dark matter”. Additionally to control the effect of the differing user profiles of each platform on the nature of the discussion, a platform by platform analysis will be performed atop the overall analysis. As a result, the recommendations made will be holistic and beneficial to the scientific community. Overall, we expect comments will be predominantly “personal opinion” and hence be low in objectivity and high in negativity. Reddit is expected to score highly in terms of relevant scientific discussion, polarity (positivity), objectivity and accuracy. Finally, we predict that the greatest source of confusion will be the mix-up between dark matter and dark energy.

3 Methodology

3.1 Platform Selection

The platforms were selected by popularity: (1) Facebook; (2) YouTube; (3) Instagram; (4) Tiktok; (5) Reddit [7]. X, formerly known as twitter, was excluded from our analysis due to the high prevalence of ”non-human” accounts as a result of the cap on tweet length causing the comment data to be unreliable [8]. A study estimated that between 4% to 23% of all Twitter accounts, out of approximately 330 million global active users, are bots [9]. Conversely, while the other social media platforms still have ”non-human accounts” their prevalence is not as high.

3.2 Comment Extraction

We divided the platforms to their various content types, if applicable, and treated these distinctions as different platforms to control for the impact of post type on the nature of discussion: TikTok (short form videos), Facebook (posts), YouTube (short form videos or ”shorts” and long form videos or ”videos”), Reddit (discussion forums), and Instagram (short form videos ”reels” and posts). Note that from now on, different content types from the same platform will be treated as a separate platform, for example Instagram reels and Instagram posts will be considered as two different platforms. Additionally several methods were applied to ensure that the data collected is unbiased and as general as possible. To achieve this, we used a Virtual Private Network (VPN) - specifically ExpressVPN - and signed out of all personal accounts. Then, after setting the location to ”Australia: Sydney”, we began by searching the keyword ”dark matter” for each platforms then selecting the top 20 links which were scientifically oriented. For Facebook, due to the increased volume of data, the search has to be refined to ”dark matter science”. This gave us a total of 140 sources. Comment extraction was completed through https://exportcomments.com/. This software automatically captured the 100 most recent comments from each links.

Platform	**N**	**L**
Facebook posts	411	188
Instagram posts	505	106
Instagram reels	598	123
Reddit	1452	360
TikTok	1524	66
Youtube videos	1598	315
Youtube shorts	1323	177
Total	**7411**	-

Table 1: *Properties of Comment Data from All Platforms, include Number of Comments (N), Average Length (characters) of Comments (L)*

In terms of data cleaning, we aimed to produce the 100 most most recent, non-duplicate and non-empty English comments for each source. By ensuring the same standard, we maintained a consistent format and volume of data across all social media platforms except for some platforms that innately have fewer comments, such as Instagram posts and reels. This consistency ensures comparisons are valid. We then conducted manual data cleaning to ensure the data fields remain consistent across all platforms, and finally compiled all comment data into a single data structure. For each platform, the final output included two files: (1) one collated file containing comment author, comment time, comment text and post URL/ID, and; (2) one file containing the post URL/ID and the corresponding post title as a dictionary. Table 1 shows the number of comments and the average character length of comments for each platform. The total number of comments for all platforms is 7411 non-duplicate, non-empty, English comments.

3.3 Comment Classification

The comments were classified to determine whether their content was scientifically relevant. The method combines a qualitative categories benchmark, large-language-model (LLM) classification, and statistical validation. A comprehensive and reliable analysis of large-scale unstructured data, especially text, is difficult and time consuming. Hence, we employed LLMs, specifically ChatGPT-4-turbo and Claude-3-Opus, to accurately categorise comments into predefined categories effectively gauging the overall scientific relevance of comments on posts related to dark matter. The implementation of LLMs for text classification has been shown to be accurate and efficient in processing and categorising large volumes of unstructured text data [10] [11].

We designed prompts to guide the LLMs to correctly classify the comments into the categories seen in Table 2, which were created based on a pilot study that factored user engagement, discussion, common themes, questions, and misconceptions about dark matter across platforms. The prompt engineering process involved focusing the AI models on the specific context of "dark matter - concept in astrophysics" and modifying the categories to include more dark matter-related quotes. This approach aligns with best practices in prompt engineering, which emphasise the importance of providing clear and context-specific instructions to LLMs [12].

To enhance focus and accuracy, meta prompts were specifically tailored to include additional context about dark matter, thereby directing the LLMs' attention and reducing incorrect classification. Each comment was processed with the following structured prompt to ensure uniformity in classification:

> You are an AI assistant trained to sort comments on videos relate to dark matter - concept in astrophysics into predefined categories. Your goal is to analyze each comment and assign the most relevant category to each set.
>
> 1. Correct theories and hypotheses of dark matter
> 2. Correct observations and evidence relate to dark matter
> 3. ... (Sequential categories as in Table 2)
>
> For each comment, first output the index of the comments follow by a semicolon, then output the index of the comment's category

This prompt will enforce the model to return a structured output that can easily be processed for further analysis.

Two strategies were adopted to avoid hallucinations and random results generated by the LLMs. First, the predicted results are aggregated into larger categories as seen in Table 2. The expectation is that this would help stabilise the predicted outcomes because of the reduced granularity of the classification task [13], allowing us to obtain a more certain result. Second, we utlised different batch sizes (15 and 30) and multiple AI models (GPT-4-turbo and Claude-3-Opus) to get 4 results for each comment. The mode of the aggregated prediction was counted for each comment and defined 'bias point', which is the number of predictions per comment (4) minus the count of the mode prediction. A one-sided t-test was performed to determine whether the bias was greater than one (allowing for some acceptable random prediction and hallucination) across all predictions.

The results showed a mean bias of 0.948 (std 0.8757), with a t-value of -5.227 and a p-value of 0.99 (with a degree of freedom of 7410), indicating that we can be confident that the predicted result is less than one at a 99% confidence level. However, further t-tests on each aggregated category revealed that some categories had a bias greater than one. To address this issue, we applied a filter to only consider predictions where the mode count was greater than or equal to 2. This filter resulted in the exclusion of 634 comments (10%) from the analysis. We then calculated the proportion of each category and then used to calculate the proportion of each aggregated category to understand the public perception of dark matter. This approach will allow for a more comprehensive and reliable analysis of the comments [14].

id	Category	id2	Aggregated Category
1	Theories and Hypotheses	a	Scientific Discussion
2	Observations and Evidence		
3	Questions and Clarifications		
4	Informative Comments		
5	Explanatory Remarks		
6	References to Additional Resources		
7	Analogies and Metaphors	b	Personal Opinion
8	Personal Opinions		
9	Philosophical or Religious Perspectives		
10	Jokes and Puns	c	Humor, Sarcasm, Trolling and Inflammatory Remarks
11	Satirical Comments		
12	Insults and Personal Attacks		
13	Controversial or Provocative Statements		
14	Unrelated Personal Anecdotes	d	Off-topic Discussions
15	Comments on Video Production or Presentation		
16	Advertisements	e	Spam and Promotional Content
17	Self-promotion		

Table 2: *Predefined Categorisation of Comments and Their Corresponding Aggregated Categories for LLMs Classification*

3.4 Sentiment Analysis for Polarity and Subjectivity

To cross-validate the accuracy of comment categorisation and to provide additional context by identifying the emotional tone of comments, we conducted a sentiment analysis- which uses natural language processing to ascertain the emotional tone or opinion expressed in a piece of text- to determine polarity and subjectivity. Sentiment analysis aided in the identification of trends or patterns of user attitudes and opinions. We used tools including Python, Pandas for data manipulation, Matplotlib for data visualisation etc., with the most important being the TextBlob library. It included a pre-defined vocabulary library that helped us to classify the sentiments of individual words and quantify the ”polarity” and ”subjectivity” of sentences. The two metrics ”polarity” and ”subjectivity” are defined as follows with example comments sampled from Facebook posts that scored the highest and lowest in either metric:

1. Polarity: A floating number between -1 to 1, where a -1 means a strong negative sentiment, 0 means neutral sentiment, and 1 means a strong positive sentiment.
    - ”*Even in 30 years of service Hubble space talescope is still the best!*” scored the highest value of polarity.
    - ”*The silly nonsense keeps on coming!!!*” scored the lowest value of polarity.
2. Subjectivity: A floating number between 0 and 1, where 1 refers to a subjective sentence with personal opinion, emotion or judgement, and 0 refers to an objective sentence, as in a factual statement.
    - ”*They call it nothing because they have nothing to work from but it could be everything, and inspiring truth that for the moment is beyond the ego and it's ability to fathom...exciting indeed. Time for the imagination!*” scored the highest value of polarity.
    - ”*Maybe ”gravity” is caused by differences in the uniformity of space? Fractal maybe?*” scored the lowest value of subjectivity.

The source of the data is a collated dataset of 7411 non-duplicate, non-empty, English comments from all social media platforms, that were collected through the aforementioned methodology in data extraction. Data cleaning was conducted first to improve the accuracy of analysis, with steps including removing special characters, numbers, and punctuation, converting to lowercase, and removing stop words (e.g., "and", "the" etc.) insignificant to the sentiments. We also applied lemmatisation, which reduced the words to their root form, in order to reduce the number of unique words and make a sentence easier to analyze. In total, we conducted 4 analyses:

1. Polarity distribution of all comments;
2. Polarity distribution by individual platforms;
3. Subjectivity distribution of all comments;
4. Subjectivity distribution by individual platforms.

3.5 Accuracy Analysis

An accuracy analysis was performed to identify the misconceptions of dark matter being discussed on social media. The accuracy analysis was a manual, qualitative process that began with randomly selecting 30 comments from each of the 7 social media platforms and their sub-content forms. Since the comments included technical discussion and jargon, an accredited professor from the University of Sydney specialising in dark matter research, was asked to provide commentary and insight into the accuracy of each of the comments. From his insights, the comments were qualitatively classified into various categories. Manual human classification was preferred ot LLMs for accuracy as we did not have the resources to train LLMs on large enough datasets to make precise statements on the subtle misconceptions and ideas regarding an already very narrow topic.

The following categories in Table 3 were devised based on comment structure (i.e. being a discussion statement or question) and then the nature of their accuracy. For discussion, this included being accurate or inaccurate and also how relevant they were to dark matter. The questions were similarly classified based on their relevancy to dark matter and whether they were well-founded or misguided. A guided question would be based on facts and other accurate, up-to-date material, whereas misguided queries would refer to unjustifiable theories and incorrect statistics. This allowed us to determine the accuracy of each platform and construct a list of common misconceptions.

	Categories	**Sub-Categories (if applicable)**
Discussion	Accurate	
	Inaccurate	MONDs
		Dark Energy
		Other
	Neutral	
	Irrelevant	
Question	Relevant	Guided
		Misguided
	Irrelevant	

Table 3: *Categories used to classify the accuracy and relevancy of discussions and question comments*

4 Results and Discussion

4.1 Comment Classification

After implementing the methodology, we created pie charts to illustrate the results.

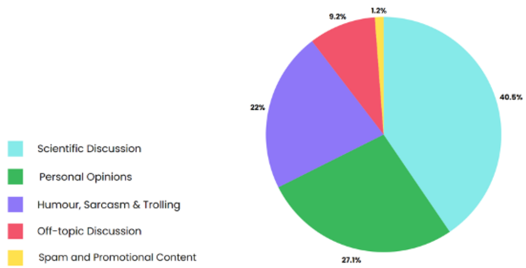

Figure 1: *Proportion of Comment categorisation in all Social Platforms*

Figure 1 indicates that scientific discussion, followed by personal opinions are the dominant categories of comments across the analysed social media platforms with the percentage of 40.5% and 27.1% respectively. This suggests a high level of engagement and interest in the topic of dark matter, accompanied by a substantial amount of humor, sarcasm, trolling and less relevant content. Spam and Promotional Content accounts for 1.2%, which is the lowest proportion.

Looking at the data across different platforms in Figure 2, comments related to scientific discussion form the largest category across all platforms, except for Youtube Short, which is outweighed by personal opinion of around 34.2%. It is apparent

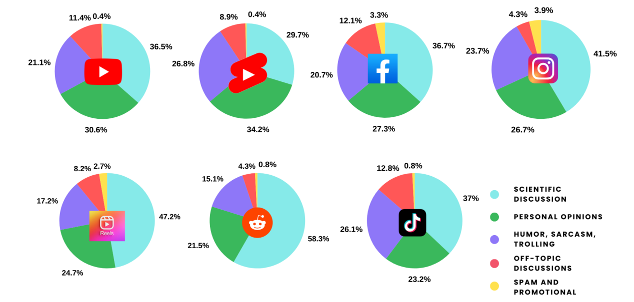

Figure 2: *Proportion of Comment categorisation Across Social Platforms*

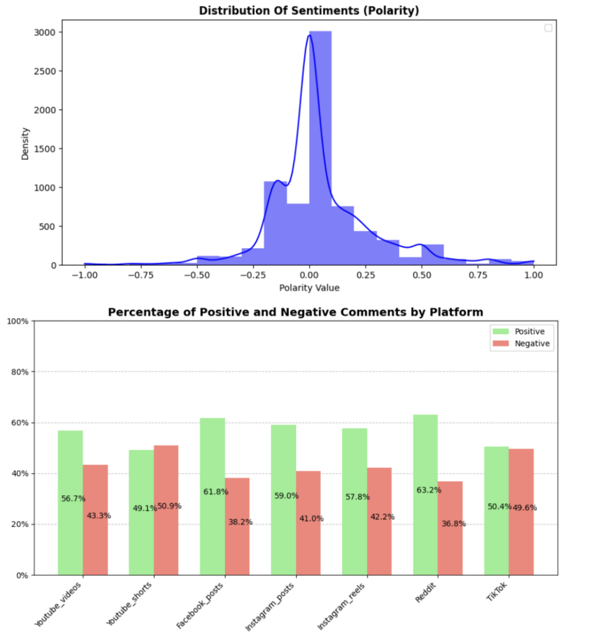

Figure 3: *Polarity distribution of all comments (up) and by platforms (down)*

that Reddit has the highest proportion of scientific discussion at roughly 58.3%. On the other hand, platforms like TikTok and Youtube shorts have the lowest levels of scientific discourse. Humour, Sarcasm, Trolling are particularly more common on Youtube Short of all platforms at about 26.8%, and for off-topic discussion is Tiktok with 12.8%. Spam and promotional content appear more frequently on Instagram at 3.9%.

Overall, we found that Reddit has the most accurate and in-depth discussion, therefore scientists can focus on posting on Reddit to generate more fruitful discussion. In Reddit, discussion is stimulated by presenting a premise alongside a question, which encourages users to give thoughtful and in-depth answers to a prompt. By combining a premise, which usually contains a context, with a question Reddit fosters a multi-faceted discussion and exploration to contribute valuable scientific insight in the community. Therefore, our results suggest that prompt-based, long-form and user-led conversations leads to much more constructive fruitful online discussion, which could be leveraged as a tool by scientists in general.

4.2 Sentiment Analysis

The polarity and subjectivity distribution of all comments were visualised using histograms with Kernel Density Estimation (KDE) to clearly identify the more populated sentiment. The x-axis is the polarity or subjectivity scores, while the y-axis is the density of comments. On the other hand, distribution comparisons by platforms were visualised using bar graphs, with the x-axis being the platforms and the y-axis being the density of comments by percentages.

In Figure 3, we examined the polarity distribution of comments. The histogram with KDE (top panel) indicated a mostly uni-modal distribution with the significant peak around 0 being the neutral sentiment, and a slight peak to its left. However, overall the right region is more populated, indicating a predominantly positive sentiment across platforms (the ratio between positive to negative comments is about 1.29). The distribution comparison by individual platform (bottom panel) indicated that positive comments greatly outweigh the negative comments in most platforms with the exception of YouTube shorts and TikTok.

A study from F. Chiossi et al [15] stated short-form videos as a feed format that represents emotional stimuli and are salient interruptions that quickly divert attention. Thus, we speculated the distribution variance in our analysis was due to the nature of the platform, as people are generally more dismissive of and more likely to express negative sentiments on short-form content which mostly targeting youthful audiences who are more likely to leave negative comments more impulsively. Nevertheless, the high negativity and objectivity in short-form video platforms does suggest reasonable scepticism and hence an inability of science communicators to convince their audiences of the validity of Dark Matter. As a result, we suggest

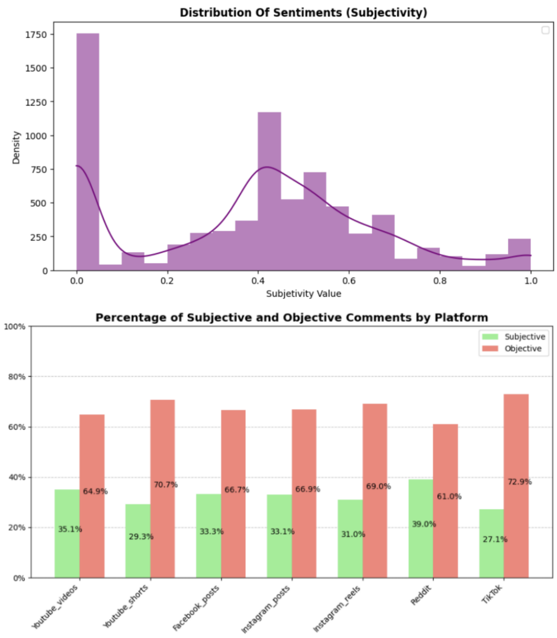

Figure 4: *Subjectivity distribution of all comments (up) and by platforms (down)*

focusing on the evidence for dark matter in short-form video platforms to reduce this scepticism. Additionally, while negative comments may be prevalent in short-form content platforms, they also present opportunities for improvement. For instance, science communicators may find it easier to identify confusion and misconceptions of the general public in these platforms to refine the quality of future content. Moreover, science communicators must also balance accessibility with accuracy when targeting youthful audiences. While brevity and simplicity are essential for engagement, accuracy and clarity should not be compromised. Strategies such as incorporating more visuals, interactive questions, and concise yet informative captions could potentially help find this balance.

In Figure 4, we examined the subjectivity distribution of comments. The histogram with KDE (top panel) indicated a bimodal distribution, with most comments being very objective (˜0), and also a large portion of ambiguous comments (˜0.5), which is expected in this context where the sentiment expressed may not easily categorized as purely subjective or objective. This echoed the purpose of including complex subcategories in our comment classification. When examining the distribution by individual platforms (bottom panel), we also observed that objective comments greatly outweigh subjective comments across all platforms. These insights could be indicative of audiences being more engaged in contents that provide facts, scientific information, tutorials etc. than contents that aimed to provoke emotions. Or, audiences of dark matter are more willing to express factual information than personal opinions in the discussion. In addition, there seemed to be little platform-specific variance in the pattern of user expression. This could suggest that the nature of the content and the topic of dark matter itself may be driving audience engagement and participation in scientific discussions, rather than specific platform algorithms or audience demographics.

Due to the limitation of time resources for this study, we validated the accuracy of the sentiment analysis mainly through human-annotated data, in which we took 5 sample comments each with (1) highest polarity; (2) highest polarity and highest subjectivity; (3) lowest polarity; (4) lowest polarity and lowest subjectivity, and manually review their given score and final sentiment classification.

4.3 Accuracy Analysis

Based on the professor's assessment, we can use Table 4 to visualise the spread of comments in the categories set out in our methodology. The most prominent misconceptions were confusion between percentages associated with dark matter, mixing up the terms dark matter, dark energy and black holes as well as a heightened estimation of modified Newtonian Dynamics as a valid alternative.

From Table 4, we can deduce that Instagram and Reddit were the platforms that contained the most accurate scientific discussion. For Reddit, this may be because the nature of the platform, a premise alongside a question, requires a greater amount of knowledge to actively want to participate in the discussion engendering lengthier and hence more accurate discussion, as seen in tables 1 and 4. For Instagram, many of the comments used in the accuracy analysis entailed short sentences that repeat the facts presented to them in the post. Therefore, Reddit is an effective platform that scientists can use if they are wanting to stimulate accurate and in-depth discussion on dark matter.

One of misconceptions were confusing similar terms such as dark matter, dark energy and black

	Categories	Sub-Categories (if needed)	Reddit	Tiktok	YouTube Videos	YouTube Shorts	Instagram Posts	Facebook	Instagram Reels	Sum of Rows
Discussion	Accurate		18	1	2	1	18	2	3	45
	Inaccurate	MONDs	1		1				1	3
		Dark Energy			4	2		2		8
		Other	5	7	3	10	4	6	10	45
	Neutral		3	6	4	1	2	5	3	24
	Irrelevant		2	1	4	7	1		5	20
Question	Relevant	Guided	1	6	6	4	1	8	5	31
		Misguided		7	6	4	2	7	3	29
	Irrelevant			2		1	2			5
Sum			30	30	30	30	30	30	30	210

Table 4: *Categorisation of Comments from Accuracy Analysis*

holes: "What is black matter n why is it called black matter?". Hence, it is recommended that scientists avoid talking about dark matter with dark energy or black holes. If these terms must be used alongside each other, then it should clearly be stated that they are distinctly different concepts.

Another common misconception was the puzzlement surrounding the percentage of the universe dark matter makes up. For example:"It covers 95% of universe matter and not 73%. Wrong info." This muddling of percentages arises from the different metrics used by scientists, one for total matter in the universe- of which dark matter is 85%- and the other is considering the entire composition of the universe wherein Dark Matter constitutes 24% [16].

There also is an over estimation of MOND as a valid alternative to the dark matter theory. Contrary to what was expected there were 5 comments that mentioned MOND out of our small sample of 210, 3 of which mentioned it inaccurately. This suggests that the public is resonating with an idea which "most astrophysicists regard the adoption of... as unjustifiable" [16]. As a result, this proves that the scientific communication of dark matter is ineffective because it promoting alternative "unjustifiable" theories as viable, this is an example: "Dark matter is a strong candidate for the missing mass, but modified Newtonian mechanics (MOND) is another possibility". The scientific community needs to reinforce the validity of accepted theories so that the public does not refer to alternatives and hence avoid confusion.

5 Conclusion

Overall 40% of discussion was scientifically relevant, with Reddit having the most scientific discussion and TikTok the least. Comments tended to be positive overall, with Reddit exhibiting the highest positivity (63.2%) and TikTok the least (50.4%), additionally, comments tended to be objective overall with Titktok having the highest objectivity and Reddit the least. As Reddit has the most accurate, scientifically relevant and positive discussion out of all platforms, mostly as expected, scientists should focus on utilising Reddit when communicating dark matter. Furthermore, TikTok and YouTube shorts evoke more negative sentiments whilst being the most objective suggesting reasonable scepticism. So in short-form videos, scientists should emphasise and prioritise the evidence behind dark matter to reduce scepticism. Additionally, the public confuses the percentages for dark matter. So when statistics are important to the explanation, the specific metric needs to be clarified to prevent perplexity. Lastly, dark matter is muddled up with dark energy and black holes. If these terms are to be mentioned together, their distinct natures have to be accentuated.

Further research should focus on a more detailed analysis of the specific social media platforms as the scope of our project only allowed us to briefly identify the trends and effectiveness of each. Nevertheless, our research will allow scientists to understand the most effective ways to utilise social media for science communication and bridge the gap between general audiences and the academic sphere.

References

[1] Conseil Européen pour la Recherche Nucléaire, *Dark matter*, 2024.

[2] Gianfranco Bertone, *The moment of truth for wimp dark matter*, 2010.

[3] F. Zwicky, "Republication of: The redshift of extragalactic nebulae," *General Relativity and Gravitation*, vol. 41, no. 1, pp. 207–224, 2009. DOI: 10.1007/s10714-008-0707-4.

[4] V. C. Rubin and J. Ford W. Kent, "Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions," vol. 159, p. 379, Feb. 1970. DOI: 10.1086/150317.

[5] K. GRIEST, "The search for the dark matter: Wimps and machos," *Annals of the New York Academy of Sciences*, vol. 688, no. 1, pp. 390–407, Jun. 1993. DOI: 10.1111/j.1749-6632.1993.tb43912.x.

[6] XENONnt Dark Matter Project, *Xenon dark matter project*, 2024.

[7] Statista. "Biggest social media platforms 2024." (2024), [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[8] IONOS Digital Guide, "Social bots – the technology behind fake news," *IONOS Digital Guide*, Mar. 2022.

[9] Z. Chen, R. S. Tanash, R. Stoll, and D. Subramanian, "Hunting malicious bots on twitter: An unsupervised approach," in *Social Informatics*, G. L. Ciampaglia, A. Mashhadi, and T. Yasseri, Eds., Cham: Springer International Publishing, 2017, pp. 501–510.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019. arXiv: 1810.04805 [cs.CL].

[11] T. B. Brown, B. Mann, N. Ryder, *et al.*, *Language models are few-shot learners*, 2020. arXiv: 2005.14165 [cs.CL].

[12] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, *Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing*, 2021. arXiv: 2107.13586 [cs.CL].

[13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, *Bertscore: Evaluating text generation with bert*, 2020. arXiv: 1904.09675 [cs.CL].

[14] H. Abburi, M. Suesserman, N. Pudota, B. Veeramani, E. Bowen, and S. Bhattacharya, *Generative ai text classification using ensemble llm approaches*, 2023. arXiv: 2309.07755 [cs.CL].

[15] F. Chiossi, L. Haliburton, C. Ou, A. M. Butz, and A. Schmidt, "Short-form videos degrade our capacity to retain intentions: Effect of context switching on prospective memory," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23, ACM, Apr. 2023. DOI: 10.1145/3544548.3580778.

[16] M. Jacquart, "cdm and mond: A debate about models or theory?" *Studies in History and Philosophy of Science Part A*, vol. 89, pp. 226–234, 2021. DOI: https://doi.org/10.1016/j.shpsa.2021.07.001.