Conspiracy Detection Beyond Text: Exploring the Feasibility of Adding Psycho-Linguistic Features to Enhance Conspiracy Detection Models

Anna George[1], Maximilian Ahrens[1], Janet Pierrehumbert[1], and Michael McMahon[1]

University of Oxford, Oxford, UK
anna.george@oii.ox.ac.uk

Abstract. Conspiracy theories pose a significant societal challenge, particularly online where their spread can be hard to detect. Robust detection models are crucial for effectively identifying these theories. In this study, we investigate incorporating emotional sentiment and moral framing features into a text-based conspiracy detection model. We hypothesize that incorporating these psycho-linguistic elements would enhance the model's performance. Our results reveal significant psycho-linguistic differences between conspiracy and non-conspiracy texts. Conspiracy texts contain higher levels of anger and are framed through the moral lens of cheating, while non-conspiracy texts contain higher levels of joy and are framed through the moral lenses of care and harm. Our model's ability to classify conspiratorial text improves after integrating emotional sentiment and moral framing into the text-based conspiracy detection model. This work demonstrates the potential value of incorporating psycho-linguistic features into text-based models to enhance conspiracy theory detection.

Keywords: Social Science · Computational Social Science · Conspiracy Theories · Classification Models · Moral Foundations Theory

1 Introduction

In today's digital landscape, both factual and misleading narratives significantly influence the dissemination of information across communities. Conspiratorial narratives, in particular, have gained prominence and pose a significant societal challenge. Their spread online is especially problematic, as it can be difficult to detect and counter. Conspiracy theories can undermine public trust, increase societal divisions, and, in some cases, lead to radicalization and extremism.

Given their potential to distort public perception and sow discord, it is urgent to establish methodologies that can effectively detect and classify these narratives as they emerge and spread online. The aim of this paper is to investigate whether incorporating psycho-linguistic features can enhance the performance of text-based conspiracy detection models. Using insights from prior research on the distinguishing linguistic features of conspiracy theories, we endeavor to

create and enhance a text-based model's classification abilities by incorporating psycho-linguistic features into the model.

Specifically, we explore the emotional sentiment and moral framing differences between conspiracy and non-conspiracy content, and evaluate whether leveraging these differences can improve automated classification of conspiratorial texts. To achieve this, we develop and compare several conspiracy detection models. We first create a text-based model for conspiracy detection, then create enhanced versions incorporating emotional sentiment and moral framing features. These psycho-linguistic features are detected using custom classifiers we develop and train on open-source datasets. By comparing the performance of these different models, we aim to determine whether, and to what extent, psycho-linguistic features can bolster conspiracy detection capabilities. Our research not only contributes to the understanding of the distinctive psycho-linguistic characteristics of conspiratorial content but also evaluates the potential of incorporating such features to enhance automated conspiracy detection methods.

Our study addresses two main research questions:

Research Question 1 (RQ1): *What are the emotional sentiment and moral framing differences between conspiracy and non-conspiracy tweets?*

Research Question 2 (RQ2): *Can the incorporation of psycho-linguistic features, specifically emotional sentiment and moral framing, improve the performance of text-based conspiracy classification models?*

By answering these questions, we hope to gain insights into the distinctive psycho-linguistic features of conspiratorial content compared to non-conspiratorial content, and evaluate the potential of incorporating such features to enhance conspiracy detection methods.

2 Literature Review

Social scientists define conspiracy theories and conspiratorial thinking in various ways. In our work, we define a conspiracy as an explanation of an event which cites an alternative explanation (e.g., alternative to official accounts) as a salient cause [7]. Lewandowsky and Cook [21] summarized the various aspects of conspiratorial thinking, stating that conspiratorial thinking often includes: contradiction, suspicion, nefarious intent [from the 'official' sources], the feeling that something must be wrong, a persecuted victim (often the person(s) engaging with conspiratorial thinking perceive themselves as the victim(s)), an immunity to evidence, and re-interpreting randomness / random events to fit the conspiracy narrative. Moreover, conspiracies can be characterized by their adversarial undertones [29] and emotive content (e.g., anger) [9].

These beliefs typically arise from a process of radicalization. Online radicalization involves a person progressively engaging with and adopting extremist ideas, leading them to extremist views and, potentially, political violence [28].

The spread of conspiracy theories is accelerated by online platforms, where individuals progressively adopt extremist content [28]. Individuals may shift away from official sources of information as they view themselves as victims being deceived by mainstream narratives [10]. Social media influencers and extremist community leaders exploit these platforms to cultivate their audience, using tactics that contribute to the appeal of alternative movements [22; 23].

The process of fringe ideas spreading to mainstream spaces, termed "normiefication" [6], occurs through multiple platforms, with some acting as "bridges" between fringe and mainstream discourse. "Bridge people" with weak ties to multiple groups facilitate the diffusion of information between communities [38]. A study on anti-vaccination conspiracy narratives on Facebook demonstrated the resilience of these narratives, showing that when conspiracy leaders were removed, other conspiracy theorists stepped in to continue spreading the information [25].

While we understand how information spreads through communities and the tactics influencers use to disseminate these narratives, little research has focused on the appeal of the narratives themselves. Given their potential harmful effects, it is crucial to develop effective methods for identifying and understanding these narratives, such as utilizing natural language processing (NLP). NLP is a way of studying language to give it meaningful computational representation [24]. Machine learning can enhance NLP methods to enable researchers to predict psychology traits [36]. A classifier is a machine learning tool that processes the words a person uses to try to decipher the underlying constructs embedded in their words. Until recently, text-based models have been limited by their inability to capture the meaning behind entire sentences and paragraphs. Now, models are able to detect more than the 'keywords' of text, and can encode entire sentences and paragraphs into a meaningful format for NLP tasks. Psychological and linguistic features can be extracted from textual information. These features tend to stem from psychological theory and have previous qualitative and experimental evidence studying their characteristics.

In our study, we classify the content features of moral frames and emotion in conspiratorial text in an attempt to enhance automated classification of conspiracy text. By analyzing these psychological and linguistic features, we hope to gain insights into the distinctive characteristics of conspiratorial content and improve detection methods. Below we review the literature on these features.

2.1 Emotion Detection

Emotions can be described as a multifaceted interplay between subjective experiences and the external world, leading to a range of outcomes: affective experiences such as feelings of happiness or sadness; cognitive processes including judgment and attention focusing; physiological adjustments like increased heart rate or sweating; and expressive behaviors, for example, smiling or frowning [19]. The theory of basic emotions identifies six primary emotions universally expressed and recognized by humans: fear, anger, joy, sadness, disgust, and surprise [8]. Other researchers propose models which organize emotions on two dimensions:

pleasure (ranging from misery to pleasure) and arousal (ranging from sleepiness to arousal). This framework allows for a fuller spectrum of emotions, such as: aroused, excited, pleased, sleepy, depressed, miserable, and distressed [33].

Emotion detection involves identifying distinct human emotion types from data sources [30]. Within textual data, emotion detection represents a specialized form of sentiment analysis that extracts fine-grained emotional states from text [1]. Analyzing the words that are being used to communicate not only gives insight into the psychological nature of the person who is expressing the words, but these words can reveal patterns of speech for entire groups of people [18; 36].

Various computational approaches are used for analyzing emotional sentiment in text. The Linguistic Inquiry and Word Count (LIWC) utilizes a lexicon-based method to gauge emotional sentiment along with other psychological states [37]. More advanced models, such as those based on transformer models, demonstrate effectiveness in capturing emotional sentiment from complex text structures, outperforming models like GPT-3 [1; 3]. In the context of conspiracy theories, prior research indicates that conspiratorial text often expresses emotions like anger and fear [9]. Incorporating emotional insights could potentially enhance the ability to detect conspiracy theories.

2.2 Moral Foundations Theory

According to Moral Foundations Theory (MFT), each person has intuitions that guide their understanding of what is moral and immoral [15]. There are many moral values that may exist and are shared between humans, but the identified and most researched moral values within MFT are: care, respect for authority, purity, fairness, and in-group loyalty. Each value has virtues and vices associated with it. Haidt et al. [15], define the moral value of care as valuing human protection, with the vice acting with cruelty (harm). Respect for authority is an obligation to submit to higher status persons, with the vice (subversion) of disobeying or showing disrespect for authority. Purity is defined as avoiding things that could be deemed disgusting or contaminating, while the vice (degradation) is being degrading or unnatural. Fairness refers to demanding justice, while the vice (cheating) involves injustice or fraud. Lastly, in-group loyalty is defined as being loyal to group affiliations (e.g., nation, family), and the vice (betrayal) is defined as betraying group affiliations [15].

The principles of MFT have been applied in various research contexts, including classification tasks. For instance, [16] leverage MFT to detect polarized concepts in online forums, particularly Reddit. By using text embeddings to project discussions into moral subspaces, the authors capture the nuanced biases in concept discussions, enhancing the detection of ideological polarization without explicit political labels. This demonstrates the potential of using Moral Foundations Theory in classification tasks, especially in the context of online discussions.

The application of MFT in online contexts is particularly relevant given the influence of sentiment on the popularity and diffusion of content. Through studying network diffusion dynamics, researchers find that Twitter (X) messages con-

taining emotional language and moral values that align with the reader are more likely to be shared widely [4]. Conversely, messages that do not align with the reader's moral values are less likely to spread online [4]. Research suggests that individuals who believe in conspiracy theories tend to express the moral values related to purity, authority, and in-group loyalty [20; 31]. This could imply that conspiracy messages are often framed with these values in mind, which may in turn facilitate their spread on social media.

Building on these insights, we hypothesize that similar emotional and moral framing differences will be present between conspiracy and non-conspiracy content in our dataset. Therefore, we pose the following research question: (RQ1) *What are the emotional sentiment and moral framing differences between conspiracy and non-conspiracy tweets?*

2.3 Enhanced Conspiracy Detection

The transmission of conspiracy theories poses a significant concern given their potential to undermine public trust and increase societal divisions [35]. Quickly, and accurately, identifying conspiracy theories allows for proactive measures to limit their spread, thereby preserving public trust and providing accurate information to the public. Although there are some examples of research combining psychological features with textual data to enhance conspiracy detection, these are limited. There are also limited instances in other fields where psychological features and textual embeddings are integrated, such as in personality prediction using social media data [5]. However, the creation of models that integrate text and psycho-linguistic features is relatively unexplored, especially in the field of conspiracy detection. The closest related work involves using psycho-linguistic features along with convolutional neural networks (CNNs) to identify individuals who propagate conspiracy theories [13].

Our research seeks to fill this gap by incorporating psycho-linguistic features into a transformer model to improve conspiracy theory detection. This model combines textual embeddings with the psycho-linguistic features of emotional sentiment and moral frames to classify texts as conspiratorial or non-conspiratorial. This integration aims to leverage the strengths of both psycho-linguistic features and text analysis techniques with the aim of creating a more accurate conspiracy detection model. Given this, we ask: (RQ2) *Can the incorporation of psycho-linguistic features, specifically emotional sentiment and moral framing, improve the performance of text-based conspiracy classification models?*

3 Methodology

3.1 Datasets

We use open-sourced datasets to train and evaluate our classification models. To work with the data, we first clean the text before training the model. During the text pre-processing, we remove URLs, symbols, and numerals from the dataset,

and convert all the text data to lowercase as these elements can introduce noise when working with text-based information.

We use the Emotion Dataset ($n = 416{,}809$) [34] to train our emotion detection model. Each tweet is annotated with theoretically derived emotions, inferred from the hashtags used within the tweet. These hashtags are then removed during the training and testing process.

To explore the moral framing of online content, we incorporate three resources: the Moral Foundations Dictionary 2.0 (n = 2,041 unique keywords, Frimer et al., 2019), Moral Foundation Twitter (X) Corpus (n = 1,386, Hoover et al., 2019), and the Moral Sentiment Reddit dataset (n = 500, George et al., 2020). These datasets contain keywords and social media posts annotated for moral expression. The Reddit dataset [12] contains Reddit posts from four sub-Reddits (r/LateStageCapitalism, r/liberal, r/Conservative, and r/The_Donald. Posts are collected in relation to four political issues (migration, abortion, climate change, and gun rights). Each Reddit post in the dataset undergoes manual coding. Annotations are made based on the virtues, vices, or absence of moral values aligned with Moral Foundations Theory.

The MFT dictionary 2.0 [11] is an enhanced version of the original MFT dictionary [14], both of which provide a list words related to each moral value. To validate this revised list's relevance to the intended moral values, [11] conduct a study involving participants from a diverse set of countries, including Spain, Egypt, Moldova, India, the United States, and Venezuela. Participants are prompted to write paragraphs that reflect specific moral values, allowing the researchers to assess the dictionary's accuracy by comparing these paragraphs to the list of keywords. The MFT Twitter (X) Corpus is an open sourced collection of hand coded moral values for several different topics [17]. Each tweet in the corpus is coded for moral values by 3 to 4 annotators. Each annotator hand codes tweets for the presence of moral values or the absence of any of the values.

We retrieve conspiracy tweets from a multi-topic conspiracy dataset (n = 3,100) [32]. The dataset is comprised of a collection of tweets specifically related to conspiracy theories about climate change, Covid-19, and Jeffrey Epstein. This dataset is hand-labeled by the researchers with a binary variable denoting the presence or absence of a conspiracy theory within the tweet. We collect the tweets for this dataset using the tweet IDs and the v.1 Twitter (X) API. We were able to obtain 1,558 tweets from the original dataset.

3.2 Emotion Detection Classifier

The base of our emotion detection classifier is the pre-trained RoBERTa language model [26]. Using transfer learning, we fine-tune RoBERTa to detect emotional sentiment by training RoBERTa on the Emotion Dataset [34], which contains 416,809 tweets labeled with one of six emotions (anger, fear, joy, love, sadness, and surprise). In the dataset, each tweet is annotated with one emotion, inferred from the hashtags used within the tweet. The dataset is divided into an 80:20 train-test split. To fine-tune the model, we use the AdamW optimizer [27]

and cross-entropy loss during training. Early stopping is used to prevent overfitting. The early stopping logic is implemented by monitoring the F1-score across epochs and terminating training if performance does not improve for 2 epochs. The best model state for the epoch with the highest f1-score is saved. The final version of the model is then run on the test dataset, and results in an overall F1 score of 0.940. The precision, recall, and F1 score values for each emotion also show the model is able to accurately detect emotions within the text. The scores for each emotion are shown in Table 1.

Table 1. Precision, Recall, and F1 Scores for the Emotion Detection Model

Emotion	Precision	Recall	F1 Score
Anger	0.95	0.95	0.95
Fear	0.87	0.94	0.91
Joy	0.92	0.99	0.96
Love	0.99	0.71	0.83
Sadness	0.98	0.98	0.98
Surprise	0.99	0.64	0.78

3.3 Moral Framing Classifier

We develop a classifier to detect the moral framing in conspiratorial posts. The classifier is developed using similar techniques to the emotional sentiment classifier, where RoBERTa is trained to classify moral framing by learning moral frames from hand coded textual data (as has been seen in previous research [12]). After combining the MFT dictionary 2.0 [11], MFT Twitter (X) Elections Corpus [17], and MFT Reddit corpus [12] (total n = 3,927), the training and evaluation data are split by a conventional 80:20 split. The cut-off score for labeling a value as present or not-present is set to the high standard of 0.90. The classifier performs well, with a high success rate of accurately classifying the data (Label ranking average precision = 0.997, evaluation loss = 0.012). Table 2 shows precision, recall, and F1 score for each value.

3.4 Conspiracy Detection Classifiers

To answer the first research question, we create a conspiracy detection classifier based on the multi-topic conspiracy dataset [32]. The model base is RoBERTa. We split the dataset into an 80:10:10 split of training, testing, and validation respectively.

All models are trained using the AdamW optimizer [27] with a learning rate of 4e-5 and the binary cross-entropy loss function. To handle the class imbalance in the training data, with more conspiracy tweets (n = 935) than non-conspiracy tweets (n = 311), we incorporate class weights into the loss function. These weights are inversely proportional to the class frequencies, assigning a higher

Table 2. Precision, Recall, and F1 Scores for Moral Framing Model

Category	Precision	Recall	F1 Score
Care	0.99	0.96	0.97
Harm	0.99	0.99	0.99
Fairness	0.99	0.99	0.99
Cheating	1.00	1.00	1.00
Loyalty	0.96	0.96	0.96
Betrayal	0.98	0.98	0.98
Authority	1.00	0.98	0.99
Subversion	0.99	1.00	1.00
Purity	0.99	1.00	1.00
Degradation	1.00	0.96	0.98

weight to the minority non-conspiracy class. This approach encourages the model to focus more on correctly identifying the underrepresented non-conspiracy class, rather than being biased towards simply predicting the majority conspiracy class by penalizing the model more for mistakes on the minority class. By penalizing mistakes on the minority class more heavily, the training process is incentivized to better represent both classes. The training process is performed over 30 epochs, with early stopping implemented if the model does not improve after 5 epochs to prevent overfitting. After each epoch, the model is evaluated on the validation set, and the best-performing model is saved.

To evaluate the impact of different features, we develop several models: a text-only baseline, models integrating text with emotion or moral frames, and a combined model with both text, emotions, and moral frames. The emotion-based detection model extends the text-only model by introducing an emotion embedding layer that transforms categorical emotion data into a continuous, fixed-size vector. These emotion vectors are concatenated with the output from the RoBERTa model's first token so that both the emotions and textual information are integrated into the model. The models which include moral frames have a similar architecture but differ in the type of data used. Unlike the categorical emotions, moral frames are represented as continuous scores ranging from 0 to 1, reflecting the strength or presence of each moral frame in the text. These scores are directly concatenated with the RoBERTa output.

Figure 1 illustrates a simplified version of the conspiracy classification process. Suppose the input tweet is "The government is lying about the COVID-19 vaccines to control us!" In the Input Layer, the tokenized textual data is processed. In the Processing Layer, the tweet is transformed into a RoBERTa embedding, converting the text into a numerical format using the RoBERTa language model. Additionally, psycho-linguistic features are converted into embeddings representing either emotions or moral frames. These embeddings are then combined in the Concatenation Layer into a single feature vector. Finally, in the Output Layer, the combined features are processed by the classifier to determines whether the tweet is classified as "Conspiracy" or "Non-Conspiracy."

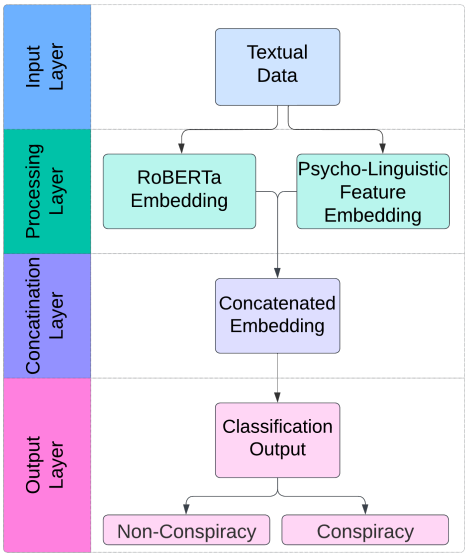


Fig. 1. Conspiracy Detection Model Architecture

4 Results

4.1 Research Question 1: What are the emotional sentiment and moral framing differences between conspiracy and non-conspiracy tweets?

Emotional Differences To classify the emotional sentiment in conspiracy tweets we use our emotion detection model. Our model outputs a probability score for each possible emotion, which represents the model's confidence that a given emotion is the most appropriate label for the text. Due to the original dataset always having an emotion label, we need to introduce the label of "no emotion". Therefore, we set a cut-off score to indicate how high the probability needs to be to label the emotion as present in the text. We choose the cut-off score based on the mean of the distribution of probability scores ($M = 0.640$). Texts are labeled with emotion with the highest probability score as long as that probability score is above 0.640, while those with scores under this threshold are considered emotionless.

Running the model on the the multi-topic conspiracy dataset [32] we apply a Chi-square test for independence to evaluate emotional differences between tweets containing or not containing conspiracies. The results indicate a significant overall variation in the distribution of emotions between the two groups ($\chi^2(6,\ n = 1{,}558) : 15.536,\ p = 0.017$). Post hoc Z-tests for individual emotions reveal that anger is significantly more common in conspiracy-related texts

($p < 0.01$), while joy is more common in non-conspiracy related texts than conspiracy related texts ($p < 0.05$). Fear, love, sadness, and surprise do not differ significantly between groups ($p > 0.05$).

Moral framing differences Similar to emotional sentiment differences, we also look at the differences in moral framing by conspiracy type in the multi-topic conspiracy dataset. To classify the moral framing in tweets we use our moral framing detection model. Our model outputs a probability score for each possible moral value, which represents the model's confidence that a given value frame is used in the text. The distribution of probability scores peaks around 0 and on the upper end towards 1, so a cut-off score of 0.90 is applied. Running the model on the the multi-topic conspiracy dataset [32] we apply a Chi-square test for independence to evaluate moral framing differences between tweets containing or not containing conspiracies.

The results indicate a significant variation in the distribution of moral frames between conspiracy and non-conspiracy tweets ($\chi^2(9, n = 1{,}558) = 34.584$, $p < 0.001$). Post hoc Z-tests for individual moral values indicate care ($p < 0.001$) and harm ($p = 0.004$) are more prevalent in non-conspiracy texts than conspiracy tweets, while Cheating is significantly more common in conspiracy-related tweets than in non-conspiracy related texts ($p < 0.001$). Other moral values such as fairness, loyalty, betrayal, authority, subversion, purity, and degradation do not show significant differences between the groups ($p > 0.05$).

4.2 Research Question 2: Can the incorporation of psycho-linguistic features, specifically emotional sentiment and moral framing, improve the performance of text-based conspiracy classification models?

To address research question 2, we build several conspiracy classification models as described in the methods section. After establishing a text-only RoBERTa conspiracy classifier as a baseline, we evaluate the impact of integrating different psycho-linguistic features related to emotions and moral framing.

The results show that incorporating all 6 emotion features boosts the F1 scores on both the validation and test sets compared to the text-only model. Given the significant differences in anger and joy between conspiracy and non-conspiracy texts observed in RQ1, we also try adding only those specific emotion features. Indeed, adding only the anger feature to the text-based model shows an improvement in the model over the text-based model alone. Similarly, the combination of joy and the text-based information improves performance over the text-based model alone, though to a lesser extent than anger.

A similar pattern is observed for moral framing features. Incorporating all 10 moral frames shows slight improvement over the text-only model. Given the significant differences in cheating, care, and harm between conspiracy and non-conspiracy texts observed in RQ1, we try adding only those specific moral frames as features. In these instances, a more substantial improvement is observed when

combining these specific moral values with the text-based model than when all 10 moral frames are included in the model. Most notably, adding both care and harm to text-based features results in the best performing model in the testing dataset, while adding all three moral frames of cheating, care, and harm results in the best model in the validation dataset. Table 3 presents the full results across these models on the validation and test datasets.

Table 3. Results of the multi-topic conspiracy detection model

Model	**F1 Score (Validation)**	**F1 Score (Test)**
Emotions	0.683	0.730
Moral Frames	0.483	0.607
Text (RoBERTa)	0.817	0.781
Text (RoBERTa) + Emotions	0.829	0.858
Text (RoBERTa) + Joy	0.836	0.856
Text (RoBERTa) + Anger	0.832	0.879
Text (RoBERTa) + Joy + Anger	0.809	0.840
Text (RoBERTa) + Moral Frames	0.823	0.800
Text (RoBERTa) + Cheating	0.814	0.826
Text (RoBERTa) + Harm	0.832	0.891
Text (RoBERTa) + Care	0.821	0.872
Text (RoBERTa) + Cheating + Care	0.814	0.892
Text (RoBERTa) + Cheating + Harm	0.829	0.889
Text (RoBERTa) + Care + Harm	0.820	**0.895**
Text (RoBERTa) + Cheating + Care + Harm	**0.843**	0.874

5 Conclusion

In this study, we investigate the psycho-linguistic differences between conspiracy and non-conspiracy texts, and if these differences can be incorporated into a conspiracy detection model to improve classification performance. The results reveal significant differences between conspiracy and non-conspiracy texts in our dataset. Consistent with previous research [9], we found that conspiracy narratives express higher anger sentiment. However, contrary to existing literature, our conspiracy texts did not exhibit higher fear sentiment [9] or emphasize moral frames like purity or loyalty [20; 31]. Instead, conspiracy texts in our dataset prominently featured moral frames related to cheating. Non-conspiracy narratives, conversely, tended to express more joy and highlight moral concerns around care and harm avoidance. These differences may be attributed due to our dataset's broader range of conspiracy topics, suggesting that emotional and moral patterns in conspiracy theories could be topic-specific. Future research could explore how these patterns vary across different conspiracy theories.

Importantly, our results demonstrate that integrating specific psycho-linguistic features related to emotions and moral framing can significantly enhance the accuracy of text-based conspiracy detection models. Incorporating features such

as anger sentiment and moral framing around cheating, harm, and care substantially improved our models' predictive performance. These findings have crucial implications for content moderation strategies on social media platforms, potentially enabling more effective identification and mitigation of harmful conspiracy theories.

However, our study has limitations. Our emotion detection relies solely on textual content, lacking access to non-verbal communicative signals that could aid in emotion recognition [2]. Additionally, while our dataset covers multiple conspiracy topics, it may not be representative of all conspiracy theories circulating online and is limited to one platform (i.e., Twitter/X). Future research should aim to validate our findings using larger and more diverse datasets, explore the effectiveness of incorporating other psycho-linguistic features, and investigate if conspiracies are expressed differently on other platforms. Furthermore, examining how emotional and moral patterns vary across different conspiracy theory topics could provide valuable insights into the nature and spread of these narratives.

Nevertheless, our work illustrates the potential of incorporating psycho-linguistic features to enhance conspiracy detection models. As conspiracy theories proliferate in the digital age, developing robust detection methods that leverage insights from multiple disciplines is crucial in combating their spread and societal impact. Our study contributes to this important goal, and our improved detection model could facilitate more timely identification and intervention, potentially mitigating the harmful effects of conspiracy theories on public discourse, trust, and decision-making. This interdisciplinary approach underscores the necessity of integrating diverse fields to address the complex challenges posed by online conspiracy theories.

Acknowledgments. The research was funded by the Alan Turing Institute and the UK Defence Science and Technology Laboratory, European Research Council (Consolidator Grant Agreement 819131), and The Dieter Schwarz Foundation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

[1] Acheampong, F.A., Nunoo-Mensah, H., Chen, W.: Transformer models for text-based emotion detection: a review of bert-based approaches. Artificial Intelligence Review **54**(8), 5789–5829 (2021)

[2] Beattie, G.W.: Language and non-verbal communication: The essential synthesis. Linguistics **19**, 1165–1183 (1981)

[3] Boitel, E., Mohasseb, A., Haig, E.: A comparative analysis of gpt-3 and bert models for text-based emotion recognition: performance, efficiency, and robustness. In: UK Workshop on Computational Intelligence. pp. 567–579. Springer (2023)

[4] Brady, W.J., Wills, J.A., Jost, J.T., Tucker, J.A., Van Bavel, J.J.: Emotion shapes the diffusion of moralized content in social networks. Proceedings of the National Academy of Sciences **114**(28), 7313–7318 (2017). https://doi.org/10.1073/pnas.1618923114

[5] Christian, H., Suhartono, D., Chowanda, A., Zamli, K.Z.: Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. Journal of Big Data **8**(1), 68 (2021)

[6] De Zeeuw, D., Hagen, S., Peeters, S., Jokubauskaite, E.: Tracing normiefication. First Monday (2020). https://doi.org/10.5210/fm.v25i11.10643

[7] Dentith, M.R.: The philosophy of conspiracy theory: Bringing the epistemology of a freighted term into the social sciences (2018)

[8] Ekman, P.: Are there basic emotions? (1992)

[9] Fong, A., Roozenbeek, J., Goldwert, D., Rathje, S., van der Linden, S.: The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on twitter. Group Processes & Intergroup Relations **24**(4), 606–623 (2021)

[10] Franks, B., Bangerter, A., Bauer, M.W.: Conspiracy theories as quasi-religious mentality: An integrated account from cognitive science, social representations theory, and frame theory. Frontiers in Psychology **4**, 424 (2013)

[11] Frimer, J.A., Boghrati, R., Haidt, J., Graham, J., Dehgani, M.: Moral foundations dictionary for linguistic analyses 2.0 (2019). https://doi.org/10.17605/OSF.IO/EZN37

[12] George, A., Bright, J.: Classifying Moral Sentiment to Measure Differences in Online Political Self-Expression. Thesis (msc), University of Oxford (2020)

[13] Giachanou, A., Ghanem, B., Rosso, P.: Detection of conspiracy propagators using psycho-linguistic characteristics. Journal of Information Science **49**(1), 3–17 (2023)

[14] Graham, J., Haidt, J., Nosek, B.: Liberals and conservatives rely on different sets of moral foundations. Journal of Personality and Social Psychology **96**(5), 1029–1046 (2009). https://doi.org/10.1037/a0015141

[15] Haidt, J., Joseph, C.: Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. Daedalus **133**(4), 55–66 (2004)
[16] Hofmann, V., Dong, X., Pierrehumbert, J.B., Schütze, H.: Modeling ideological salience and framing in polarized online groups with graph neural networks and structured sparsity. arXiv preprint arXiv:2104.08829 (2021)
[17] Hoover, J., et al.: Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. Social Psychological and Personality Science (2019), dIO: 1948550619876629
[18] Khalid, O., Srinivasan, P.: Style matters! investigating linguistic style in online communities. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 14, pp. 360–369. AAAI (2020)
[19] Kleinginna Jr, P.R., Kleinginna, A.M.: A categorized list of emotion definitions, with suggestions for a consensual definition. Motivation and emotion **5**(4), 345–379 (1981)
[20] Leone, L., Giacomantonio, M., Lauriola, M.: Moral foundations, worldviews, moral absolutism and belief in conspiracy theories. International journal of psychology **54**(2), 197–204 (2019)
[21] Lewandowsky, S., Cook, J.: The conspiracy theory handbook (2020)
[22] Lewis, B.: Alternative influence. Tech. rep., Data & Society; Data & Society Research Institute (2018)
[23] Lewis, R.: "this is what the news won't show you": YouTube creators and the reactionary politics of micro-celebrity. Television & New Media **21**(2), 201–217 (2020). https://doi.org/10.1177/1527476419879919
[24] Liddy, E.: Natural language processing. In: Encyclopedia of Library and Information Science. Marcel Decker, Inc., New York, 2 edn. (2001)
[25] Ligot, D., Tayco, F.C., Toledo, M., Nazareno, C., Brennan-Rieder, D.: Infodemiology: Computational methodologies for quantifying and visualizing key characteristics of the COVID-19 infodemic. SSRN Electronic Journal (2021). https://doi.org/10.2139/ssrn.3771695
[26] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
[27] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
[28] Marwick, A.E., Furl, K.: Taking the redpill: Talking about extremism. AoIR Selected Papers of Internet Research (2021)
[29] Moffitt, J.D., King, C.: Hunting conspiracy theories during the COVID-19 pandemic. Social Media+ Society **7**(3) (2021)
[30] Nandwani, P., Verma, R.: A review on sentiment analysis and emotion detection from text. Social network analysis and mining **11**(1), 81 (2021)
[31] Nejat, P., Heirani-Tabas, A., Nazarpour, M.M.: Moral foundations are better predictors of belief in covid-19 conspiracy theories than the big five personality traits. Frontiers in Psychology **14**, 1201695 (2023)
[32] Phillips, S.C., Ng, L.H.X., Carley, K.M.: Hoaxes and hidden agendas: A twitter conspiracy theory dataset: Data paper. In: Companion Proceedings of the Web Conference 2022. pp. 876–880 (April 2022)

[33] Russell, J.A.: A circumplex model of affect. Journal of personality and social psychology **39**(6), 1161 (1980)
[34] Saravia, E., Liu, H.C.T., Huang, Y.H., Wu, J., Chen, Y.S.: CARER: Contextualized affect representations for emotion recognition. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3687–3697. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1404
[35] Sunstein, C.R., Vermeule, A.: Conspiracy theories: Causes and cures. Journal of political philosophy **17**(2), 202–227 (2009)
[36] Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: Liwc and computerized text analysis methods. Journal of Language and Social Psychology **29**(1), 24–54 (2010)
[37] Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology **29**(1), 24–54 (2010)
[38] Zhao, J., Wu, J., Xu, K.: Weak ties: Subtle role of information diffusion in online social networks. Physical Review E **82** (2010)