Unveiling Online Conspiracy Theorists: a Text-Based Approach and Characterization

Alessandra Recordare[1,*], Guglielmo Cola[1], Tiziano Fagni[1] and Maurizio Tesconi[1]

[1]Insitute of Informatics and Telematics (IIT), National Research Council (CNR), Via G. Moruzzi 1, 56124, Pisa, Italy

Abstract

In today's digital landscape, the proliferation of conspiracy theories within the disinformation ecosystem of online platforms represents a growing concern. This paper delves into the complexities of this phenomenon. We conducted a comprehensive analysis of two distinct X (formerly known as Twitter) datasets: one comprising users with conspiracy theorizing patterns and another made of users lacking such tendencies and thus serving as a control group. The distinguishing factors between these two groups are explored across three dimensions: emotions, idioms, and linguistic features. Our findings reveal marked differences in the lexicon and language adopted by conspiracy theorists with respect to other users. We developed a machine learning classifier capable of identifying users who propagate conspiracy theories based on a rich set of 871 features. The results demonstrate high accuracy, with an average F1 score of 0.88. Moreover, this paper unveils the most discriminating characteristics that define conspiracy theory propagators.

Keywords

conspiracy theorist, disinformation, fake news, zero-shot learning, social media

1. Introduction

In the era of social networks, where the proliferation of misinformation and conspiracy theories has become a growing concern, the need to identify users responsible for creating misleading content has become imperative. In addressing disinformation, it is crucial to consider the role of social networks, as they have been shown to act as significant amplifiers [1]. That is why examining the spread of disinformation within social networks has become an area of growing research interest [2, 3, 4]. In particular, in the aftermath of the COVID-19 pandemic, there has been an increased focus on the study and understanding of conspiracy theories in general. This interest stems from the awareness of the significant impact that these theories can have on public health, social cohesion, and the dissemination of accurate information. In response to this challenge, this study aims to provide a contribution by explaining an approach to identifying and profiling individuals who promote conspiracy theories on social media [5].

This study builds upon prior research [6] where a technique was introduced to collect two datasets: one consisting of apparent conspiracy theorists and the other of generic users, all sourced from X (Twitter). Additionally, in that work a classification study was conducted to differentiate between conspiracy and generic users, using a combination of psycholinguistic features and platform-specific profile characteristics, including Following Count, Follower Count, Bio Sentences, Retweet Ratio, and more. In our research, we seek to characterize conspiracy theorists solely based on their writing style, moving away from dependencies on social network-related features. We explore three distinct categories of features: emotions, idioms, and linguistic attributes. Moreover, we aim to identify the specific features that prove to be crucial in making this distinction. Given the definition of "conspiracy user" as someone who believes in conspiracy theories (conspiracy theorist), our research questions are:

RQ1 – Is it possible to identify a conspiracy user through text alone?

RQ2 – What are the features that differentiate a conspiracy user from a generic user?

ITASEC 2024: Italian Conference on CyberSecurity, April 08–11, 2024, Salerno, Italy

*Corresponding author.

alessandra.recordare@iit.cnr.it (A. Recordare); guglielmo.cola@iit.cnr.it (G. Cola); tiziano.fagni@iit.cnr.it (T. Fagni); maurizio.tesconi@iit.cnr.it (M. Tesconi)

Our investigation involved training various classifiers using three distinct classes of text-based features, ensuring that our insights could be applied independently of the specific social networking platform. The results show that the two groups of users exhibit divergent writing styles, underscoring distinct attitudes. Moreover, we identified which types of features are most effective in revealing the tendency to adhere to conspiracy theories.

The paper is organized as follows. In the following Section we briefly present some of the most relevant studies in the field of fake news and conspiracy theories detection. In Section 3, we describe the X dataset and the preprocessing steps required for our analysis. Next, in Section 4.1, we show and describe the adopted features. Section 5 shows the analysis we performed on the dataset. Finally, in Section 6, we summarize our findings and suggest avenues for future research.

2. Related work

Social media have greatly facilitated the dissemination of unverified information and misleading content [1, 3, 7]. There has been extensive research on fake news detection to enable the analysis of fake news spreaders. A relevant example is the study in [8], which revealed that there are characteristics (most dependent on social media) that differ between users who share fake news and those who share real news. Giachanou et al. in [9] proposed a system that exploits a set of psycholinguistic characteristics and personality traits inferred by users to discriminate between potential spreaders of fake news and fact-checkers.

Recent years have seen an increasing focus on the study of conspiracy theories among fake news and disinformation, especially in response to global events such as the COVID-19 pandemic [1, 10]. Alternative explanations for historical or ongoing events, which claim that individuals or groups with malevolent intentions are involved in occult conspiracies, have infiltrated online communication, popular culture, political discourse, and various other areas [11]. Researchers are studying how these theories spread across different social platforms, analyzing the mechanisms that lead to their adoption by individuals and trying to identify and characterize conspiracy users in different ways [12, 13, 14, 15, 16, 17]. While conspiracy theories have recently been associated with vaccines, their scope extends to several other realms. For example, Marcellino et al. [5] collected and analyzed online discussions related to four specific conspiracy theories. Klein et al. [18] examined users posting a variety of conspiracy theories on Reddit, analyzing differences in the language used by conspiracy theorists compared to other users. The work of Fong et al. [19] analyzed conspiracy theories posted by influencer users on Twitter. Bessi et al. [20] examined the differences between Facebook users who adhere to conspiracy theories and those who do not, characterizing the personalities of the two groups.

Many of these studies have suggested the possibility of distinguishing the two user groups, yet they lack detailed insights into the extent of this differentiation. Our aim is therefore to distinguish between "conspiracy users" and other users on social media and to give them a characterization. Our work differs from those mentioned above in that we determine whether a user is a conspiracy user by using only the text of posted tweets, independently of other dynamics of the social networking platform. In this context, a relevant study is presented in [21], where a classifier for conspiracy users is described. However, unlike their approach, we do not seek a distinction between users who support conspiracy theories and users who refute them, but instead compare apparent conspiracy theorists with generic users discussing the same topics. Additionally, while their focus was on a narrow range of conspiracies, our study considers a broader set of conspiracy theories.

3. Dataset description

In this section, we present a detailed account of the initial dataset sourced from [6] as well as the specific preprocessing steps that were executed to adapt the dataset to the objectives of our research.

This initial dataset includes two distinct sets, each consisting of 7,394 X users. The first set, the "conspiracy group", comprises users identified as conspiracy theorists. The second set, the "control

group", includes users not exhibiting apparent conspiracy theory patterns. Conspiracy users were identified by analyzing likes and follows of well-known conspiracy pages or accounts. Instead, the control group consists of users who neither explicitly liked nor followed such content, but still engaged in discussions on the same controversial topics as the conspiracy group and were created around the same time. For each user, the last 3,200 tweets were collected, as of June 13, 2022. The dataset is publicly available[1].

To ensure the dataset's relevance and reliability for our research objectives, a series of preprocessing steps were undertaken:

- **Removal of Retweets**: To enhance the dataset's suitability for profiling users, we chose to exclude retweets. Retweets, being reposts of other's content, introduce redundancy. By excluding them, we ensured that the dataset primarily consists of original content, aligning with our goal of characterizing users based on their own tweets.
- **Language Filter**: Our analysis focused exclusively on tweets composed in the English language. Implementing this filter was crucial for the subsequent phases of our research and ensured linguistic coherence in our dataset.
- **User Tweet Count Threshold**: To ensure the inclusion of users who have a sufficient presence on X, we implemented a per-user tweet count threshold. Specifically, we excluded users with fewer than 10 tweets within the data collection period. This helped improve the accuracy and reliability of the user profiling we aimed to achieve.
- **Selection of Latest 100 Tweets per User**: Obtaining a large number of tweets from a single user is often challenging in practice. To address this, we focused our analysis on the most recent 100 tweets for each user. This approach reflects more closely real-world scenarios, where several users do not have a high volume of tweet activity.

These preprocessing steps transformed the dataset into a more suitable form for our research objectives. As a result, our dataset contained 547,724 tweets from conspiracy users and 592,927 tweets from the control group, posted by a total of 14,568 users. We then balanced the dataset using a Random Undersampling technique, achieving a total of 7,210 conspiracy users and 7,210 control group users.

4. Method

The objective of our study is to characterize conspiracy users through a series of steps: identifying suitable features that are dependent solely on the text of the tweet and are not influenced by the platform, conducting a classification task to distinguish between the two groups, and analyzing the most significant features using feature importance metrics. This analysis aims to discern the stylistic differences between the two user groups while ensuring the exclusion of platform-related factors.

In this section we present the features used to characterize users based on their tweets as well as the classification methods employed to discriminate between conspiracy theorists and other users.

4.1. Features

We opted to employ three distinct feature groups, all centered around the text content of each individual tweet:

1. **Emotions**: We included this feature group to partially implement a sentiment analysis on the dataset. The emotions we have chosen are *Anger*, *Fear*, *Joy*, *Sadness*, *Disgust*, *Surprise*, *Anticipation*, and *Trust*. These emotions align with Robert Plutchik's model of basic emotions, which is widely recognized in the field of psychology [22]. To assess the emotional content of each tweet, we employed zero-shot learning techniques. Specifically, we used the `facebook/bart-large-mnli`

[1] https://zenodo.org/records/8239530

model available on Hugging Face [2] as a sentiment classifier. This pre-trained model provides a score on a scale from 0 to 1, where a score of 0 indicates no agreement between the emotion and the tweet, while values approaching 1 indicate a strong agreement between them. The facebook/bart-large-mnli model is well known for its accuracy in discerning emotional content in text data [23]. This agreement calculated between the emotion and the tweet will be the feature used for our work.

2. **Idioms of conspiracy theorists**: 44 sentences were generated by chatGPT-3.5 using the following prompt:

```
What are the typical idioms of a conspiracy theorist?
Some sayings that come to mind are:
> - think/reason/. . . with your head
> - they won't tell you any of this
> - they don't tell us
> - nobody talks about it
> - wake up!
> - strong powers
> - they make fun of us
> - that's enough
Do you know any other interesting ones?
```

   These idioms are detailed in Table 1 and they aim to represent the typical language used by conspiracy theorists on social media. The agreement between tweets and idioms was assessed using the zero-shot learning capability of the facebook/bart-large-mnli model, similarly to the method used for emotions. The agreement score, ranging from 0 to 1, was utilized as a feature for subsequent analyses.

3. **Linguistic features**: We have identified five sets of linguistic features for a total of 72: *lexical* (e.g., num_words),*syntactical* (e.g. num_sentences), *semantic* (e.g., num_named_entities), *structural* (e.g., avg_sentence_length), and *subject-specific* features (e.g., flesch_reading_ease). The full list is reported in Table 2.

Table 3 shows the number of features per each group.

After selecting these features to characterize users, we decided to aggregate tweets by individual users. This transformation resulted in a dataset where each row represents an individual user, rather than an individual tweet. To achieve this, we computed 7 descriptive statistics, namely the mean, median, standard deviation, minimum, maximum, lower quartile, and uppr quartile for the values associated with the tweets of the same user. For instance, when considering the feature 'num_sentences', the statistics were computed by aggregating values of 'num_sentences' across all tweets authored by an individual user. As a result, each feature was expanded into 7 distinct statistical measures. This aggregation process provided us with a more holistic perspective on the characteristics of each user, facilitating the summarization of tweet-level attributes into user-level attributes. After this step, our final dataset comprised 14,420 rows (users) and 868 features.

4.2. Classification

The dataset was split into training (85%) and test (15%) sets. Classification was conducted using either the three groups of features individually or a combined set incorporating all of them. We evaluated a variety of classifiers, including Logistic Regression, K-Nearest Neighbours (K-NN), Naive Bayes, Support Vector Machine (SVM), Decision Trees, Random Forest, Gradient Boosting, such as XGBoost and LightGBM, Quadratic Discriminant Analysis (QDA), Multilayer Perceptron (MLP), Ridge Classifier, and Linear Discriminant Analysis (LDA). For each classification algorithm, stratified k-fold cross-validation was utilized on the training set to fine-tune the parameters.

[2] https://huggingface.co/facebook/bart-large-mnli

Idioms	
Behind closed doors	They want to keep us in the dark
Don't let them catch you	They will not tell you anything about this
Don't let the cat out of the bag	They're cooking up something nefarious
Follow the money	They're out to get us
It's a cover-up	They're planning something behind our backs
It's a deep state conspiracy	They're plotting something sinister
It's all part of the plan	They're trying to cover up their tracks
Nobody talks about it	They're trying to distract us from the real issue
Now enough!	They're trying to divide us
Pulling the strings	They're trying to silence us
Pulling the wool over our eyes	Thinking with your head
Question everything	Trust no one
Strong powers	Wake up!
The conspiracy runs deep	Watch your back
The enemy is among us	We have to be prepared
The truth is hidden	We have to stay one step ahead of them
The truth is out there	We have to stick together
The truth is suppressed	We have to watch our backs
The truth will set us free	We need to be careful who we trust
They don't tell us	We need to dig deeper and uncover the truth
They don't want us to know the truth	We need to stay one step ahead of them
They make fun of us	We need to uncover their secrets

Table 1
List of idioms used among conspiracy theorists

5. Result and Discussion

We first report the results achieved in recognizing conspiracy users from the text contained in their tweets (RQ1). Subsequently, we explore the feature importance within the three groups defined in Section 4.1, in order to unveil the key features that characterize conspiracy theorists (RQ2).

5.1. Conspiracy users classification

In the classification task, as mentioned above, we evaluated various classifiers using a single group of features (emotions, idioms, or linguistic) or all of them combined. For emotions, the best performance was achieved with Logistic Regression. For idioms, the best results were obtained through Logistic Regression, Ridge Classifier, and Linear Discriminant Analysis (LDA). On the other hand, for linguistic features and for the combined features, the best performances were achieved using the Light Gradient Boosting Machine (LGBM) algorithm. Classification results are shown in Table 4.

5.2. Feature importance

From these results, it is apparent that the feature group which excels at distinguishing between conspiracy users and control group is the set of linguistic features. Figure 1 shows the 20 most important features for discriminating between control group users (on the left) and conspiracy users (on the right). Blue points represent low feature values, while red points indicate high values. The SHAP value (the distance from the central vertical axis) indicates the importance of that feature for classification. The analysis of the 20 most crucial features for classification shows that the top 10, in terms of importance, originate from the linguistic feature group, with the remaining 10 linked to idioms. Notably, none of the top 20 features are related to emotions, suggesting that emotional features have relatively limited discriminatory power between the two user groups. The most discriminative feature is *mean(num_coord_clauses)*, showing lower values for conspiracy users, followed by

Class	Features	Description
Lexical	num_words; num_unique_words; num_chars; num_unique_chars; avg_word_length; num_stop_words; num_punct; num_digits; num_upper_case_words; num_lower_case_words; num_title_case_words; num_proper_nouns; num_nouns; num_verbs; num_adjectives; num_adverbs; num_pronouns; num_named_entities; num_noun_chunks; num_exclamation_marks; num_question_marks; num_spaces	Word-level characteristics and properties of text. They include various measurements related to the vocabulary and composition of words within a given text.
Syntactical	nominal_forms; voc_rich; num_sentences; avg_num_words_per_sentence; num_noun_phrases; num_verb_phrases; num_adj_phrases; num_adv_phrases; num_prep_phrases; num_coord_conj; num_subord_conj; num_coord_clauses; num_subord_clauses; punctuation_freq; num_capitalized_sentences; num_caps_word_freq; num_participial; num_present_tense; num_complementation; num_relative_clause	Grammatical structure and syntax of sentences within a text. They capture the organization and relationships of words and phrases in terms of syntactic rules.
Semantic	num_personal_pronouns; num_impersonal_pronouns; num_possessive_pronouns; num_reflexive_pronouns; num_reciprocal_pronouns; num_quantifiers; num_determiners; num_prepositions; num_aux_verbs; num_modal_verbs; num_negations; num_synonym; num_antonymy; 1st_person_pronouns; 2nd_person_pronouns; num_passive_verbs	Meaning and interpretation of words and phrases within a text. They capture the underlying semantics and context of language.
Structural	avg_sentence_length; avg_word_length; avg_noun_phrases_per_sentence; avg_verbs_per_sentence; proper_noun_ratio	Overall organization and composition of the text at a higher level, such as sentence and paragraph structure. They provide insights into the textual coherence and complexity.
Subject-specific	flesch_reading_ease; smog_index; flesch_kincaid_grade; coleman_liau_index; automated_readability_index; dale_chall_readability_score; difficult_words; linsear_write_formula; gunning_fog	Specialized indicators relevant to specific domains or topics within the text.

Table 2
List of linguistic features divided by class

Emotions	Idioms	Linguistic Features				
		Lexical	Syntactical	Semantic	Structural	Subject-specific
8	44	22	20	16	5	9

Table 3
Number of features used per group

mean(num_reflexive_pronouons) and *mean(num_possessive_pronouons)*, both showing higher values for conspiracy users.

Figure 2 depicts the variation of the F1 score as a function of the number of features employed in the classification, arranged according to their order of importance. We can see that by utilizing the first 30 features, the maximum F1 score is achieved, and notably, even with just the first 14 features, an F1 score of 0.85 is attained.

In the following subsections, we provide a detailed analysis on the relevance of each group of features in recognizing conspiracy users, directly addressing our second research question (RQ2).

5.3. Emotions

We conducted an in-depth analysis of emotion-based feature importance in the LGBM classifier and observed that the most prominent distinguishing emotion between the two user groups is “disgust”, followed by “joy”, “sadness”, and “anticipation”. Figure 3 shows the 20 most important emotional features for discriminating between control group users and conspiracy users.

Classifier	Emotion			Idioms			Linguistic features			All features		
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Logistic regression	**0.74**	**0.81**	**0.77**	**0.80**	**0.85**	**0.82**	0.79	0.83	0.81	0.83	0.85	0.84
K-NN	0.70	0.73	0.72	0.77	0.69	0.73	0.70	0.76	0.73	0.76	0.75	0.75
Naive Bayes	0.61	0.95	0.74	0.65	0.92	0.76	0.54	0.98	0.69	0.60	0.96	0.73
SVM	0.75	0.78	0.76	0.78	0.69	0.73	0.76	0.78	0.77	0.82	0.85	0.83
MLP	0.73	0.74	0.73	0.81	0.80	0.81	0.80	0.78	0.79	0.85	0.84	0.85
Ridge Classifier	**0.73**	**0.82**	**0.77**	**0.80**	**0.85**	**0.82**	0.78	0.83	0.80	0.81	0.87	0.84
LDA	**0.73**	**0.82**	**0.77**	**0.80**	**0.85**	**0.82**	0.78	0.83	0.80	0.81	0.87	0.84
DT	0.68	0.67	0.67	0.70	0.70	0.70	0.70	0.70	0.70	0.72	0.70	0.71
RF	0.73	0.79	0.76	0.76	0.84	0.79	0.77	0.77	0.77	0.78	0.84	0.81
XGBoost	0.73	0.77	0.75	0.78	0.85	0.81	0.83	0.87	0.85	0.85	0.88	0.86
LightGBM	0.74	0.79	0.76	0.78	0.85	0.81	**0.84**	**0.87**	**0.86**	**0.86**	**0.89**	**0.87**
Mean	0.703	0.802	0.752	0.756	0.818	0.780	0.731	0.751	0.716	0.766	0.779	0.752

Table 4
Precision, Recall and F1 score on test sets

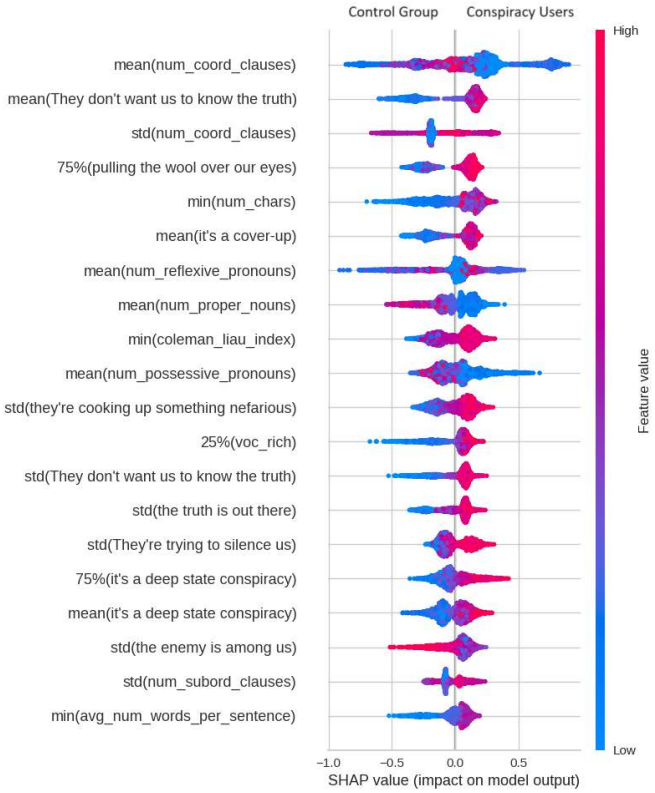

Figure 1: SHAP values for the 20 most important features considering all features

For the "disgust" emotion, the average, median, standard deviation, and 75th percentile values were significantly higher for conspiracy users. In the control group, the mean and seventy-fifth percentile of the "joy" emotion exhibited higher values. As for the "sadness" emotion, conspiracy users showed higher mean and 75th percentile values compared to the control group. Interestingly, for the "anger" emotion, both the mean and median were higher in the control group.

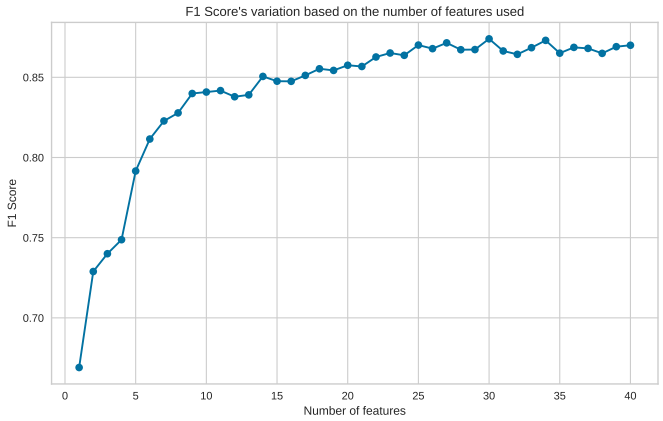

Figure 2: F1 score based on the number of features used for classification (ordered by feature importance) considering all features

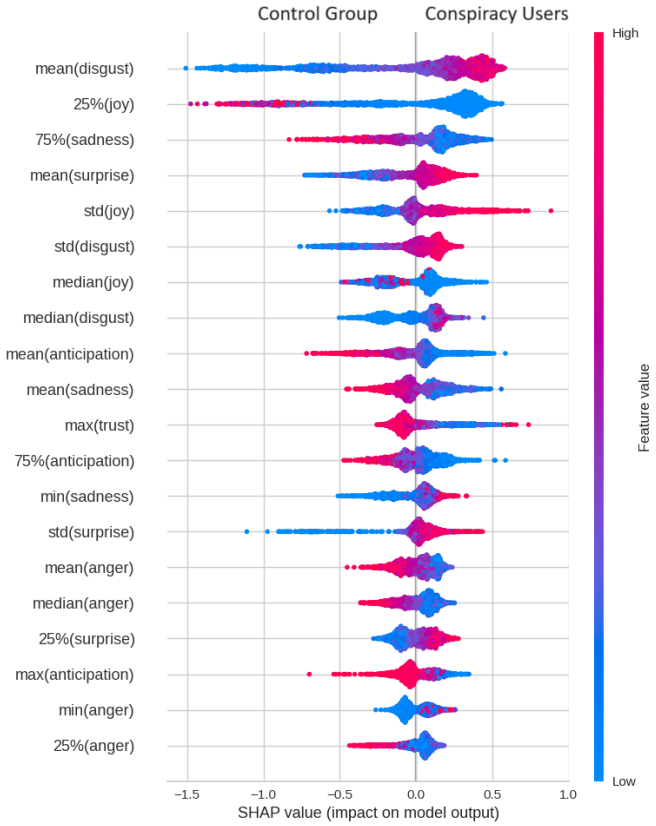

Figure 3: SHAP values for the 20 most important features in the emotions set

5.4. Idioms of conspiracy theorists

Figure 4 is a heatmap showing the average values of several descriptive statistics for the majority of idioms in our analysis, divided by user group (control group and conspiracy users). The average of each descriptive statistic was computed among all the users in a group. Cells with higher values tend towards yellow, whereas lower values are represented by violet blue cells. We excluded the least discriminating idioms and statistics for better readability.

Most of the idioms identified by ChatGPT tend to align more closely, on average, with tweets from conspiracy theorists, except for *We have to stick together*, *Strong powers*, *It's all part of the plan*, *The*

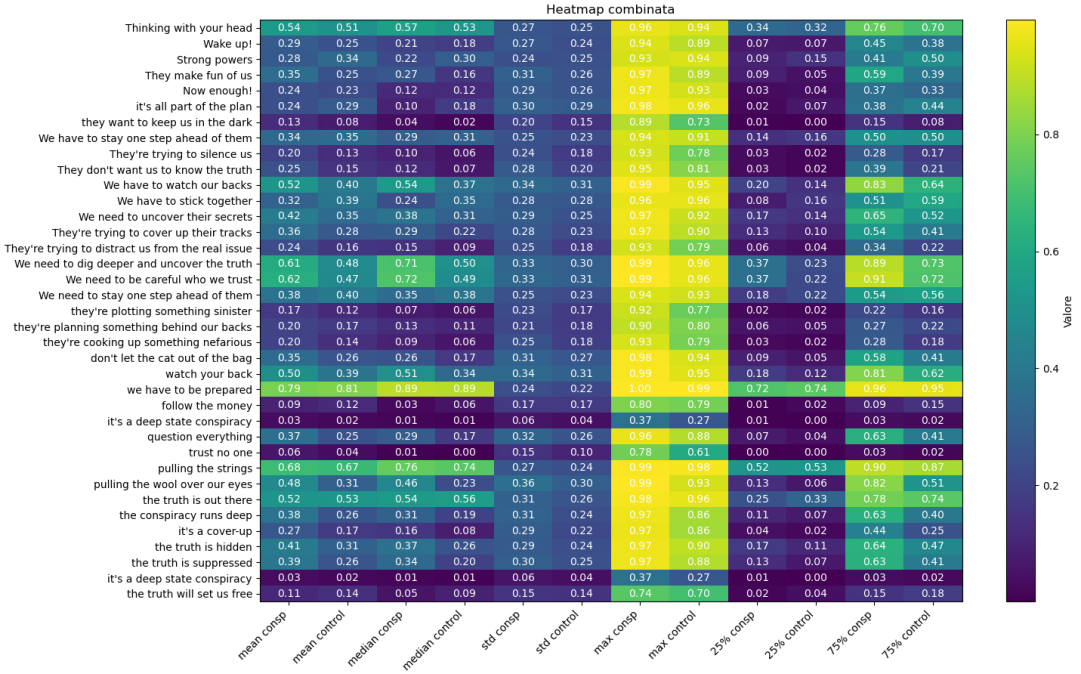

Figure 4: Descriptive statistics relative to the analyzed idioms, for control group users and conspiracy users

truth will set us free, and *Follow the money*, which align more with tweets from the control group. Some idioms exhibit a strong agreement with both user groups, like *We have to be prepared*, while others show little agreement for either group, such as *Trust no one*, *They're plotting something nefarious*, and *The enemy is among us*. The average standard deviations are consistently higher for conspiracy theorists, suggesting that this group has more diverse data among themselves. There are substantial differences in the averages of the 75th percentiles, for example, in phrases like *Pulling the wool over our eyes*, *Question everything*, and *The conspiracy runs deep*, where quite higher values are noted for conspiracy users. This indicates that agreement values for these idioms tend to be higher for this class of users.

5.5. Linguistic features

As previously mentioned, linguistic features have proven to be the most effective in classifying conspiracy and control group users. To further explore this, we divided these features into five groups: lexical, syntactical, semantic, structural, and subject-specific features. Our goal was to ascertain which of these groups contributed most significantly to the differentiation between the two user classes. Regarding their utility for classification, we found that semantic features ranked the highest, followed by syntactical, lexical, subject-specific, and, lastly, structural features. This observation is corroborated by Figure 5, which illustrates that the top 20 features contributing to classification predominantly belong to the semantic, syntactical, or lexical categories.

Among the most significant semantic features are the average count of reflexive pronouns, the average count of possessive pronouns (in both cases, the number of pronouns mentioned is higher for conspiracy users), and the average count of named entities (conspiracy users tend to mention fewer entities). As for syntactical features, the mean and standard deviation of the number of coordinating clauses, along with the standard deviation of the number of subordinate clauses, were identified as the most important. Both of these features exhibited higher values among conspiracy theorists. Among the prominent lexical features, the mean count of digits (higher for conspiracy users) and the standard deviation of title case word count (also higher for conspiracy users) were found to be the most influential. Other noteworthy features include the higher count of question marks among conspiracy users, as well as vocabulary richness, indicating a more sophisticated word choice in their tweets. Furthermore,

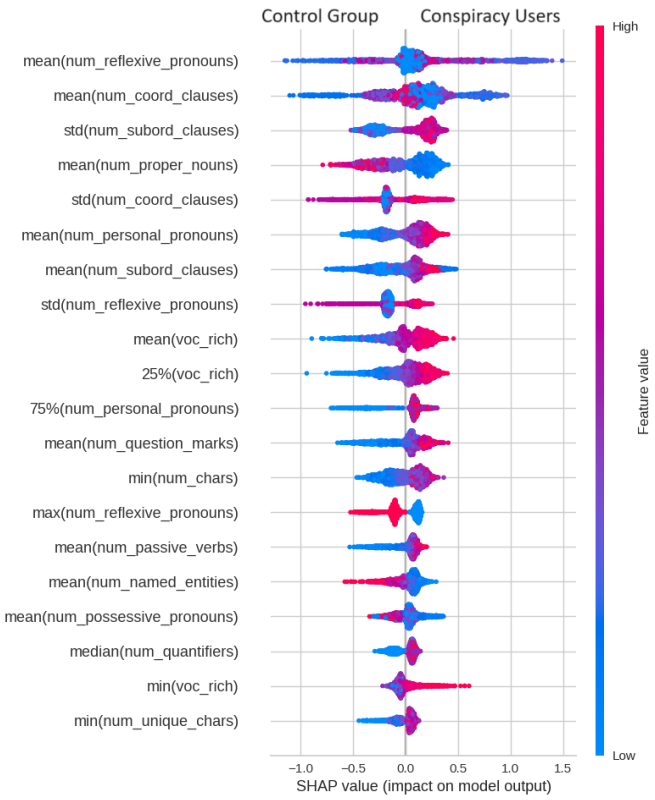

Figure 5: SHAP values for the 20 most important features in the linguistic set

all readability indices suggest a greater level of reading difficulty (and therefore lower readability) in tweets of conspiracy users. This is likely attributed to the usage of acronyms or hashtags typical of the movement they support.

6. Conclusions and future work

In this study, we introduced a method for profiling users who endorse conspiracy theories, focusing specifically on characterizing their writing style. This characterization was achieved by analyzing textual content of tweets, intentionally excluding platform-dependent metrics such as likes, retweets, and comments.

We selected a dataset consisting of 14,420 users, evenly split between two categories: 7,210 conspiracy users and 7,210 control group users, who did not exhibit explicit conspiracy theory behavior patterns. For each user, we analyzed between 10 to 100 of their most recent tweets, calculating scores based solely on textual content. These scores were subsequently aggregated for each user, using statistical measures like mean and median to capture the essence of each user's textual patterns. We also implemented and tested classification algorithms, with the Light Gradient Boosting Machine classifier yielding the most promising results. This classifier enabled us to effectively differentiate between conspiracy and control users, achieving an F1 score of 0.87.

Responding to RQ1, this research has shown that users can be categorized based solely on the characteristics of their writing style. Furthermore, in response to RQ2, this study identified specific linguistic traits that can be considered characteristic of conspiracy theorists, thus shedding light on the distinct markers of this group within the digital landscape. We found that the features that best characterize conspiracy users from the control group are linguistic features, in particular the number of coordinate clauses, the number of possessive and reflexive pronouns. Our work shows that conspiracy theorists use fewer coordinate clauses than the control group but more reflexive and possessive pronouns,

use more digits, name fewer entities, use a richer vocabulary and have worse readability. Regarding sentiment analysis, the tweets from conspiracy users show a higher agreement with disgust and sadness, while the tweets of the control group are more akin to joy and anger. Considering the set of conspiracy idioms generated via chat-GPT, it turns out that most of them have a higher agreement with conspiracy users.

In future work, we plan to extend our classifier's application to other platforms like Telegram. This will help assess the model's generalizability and robustness across diverse social media. Furthermore, we plan to improve the accuracy of our model by incorporating a wider range of text-only features, enhancing our understanding of user behavior and the overall ability of recognizing conspiracy theorists. Additionally, we are interested in exploring different time windows to capture evolving trends and emerging patterns in the propagation of conspiracy theories. Lastly, an interesting avenue for future research is examining the implications of our findings on disinformation mitigation strategies. This could lead to more effective methods to counter the spread of disinformation and promote digital literacy.

Acknowledgments

We acknowledge the support provided by project SoBigData.it, which receives funding from European Union – NextGenerationEU – National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) – Project: "SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics" – Prot. IR0000013 – Avviso n. 3264 del 28/12/2021. This work is also supported by project SERICS (PE00000014) under the NRRP MUR program funded by the EU – NGEU.

References

[1] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, et al., The covid-19 social media infodemic, Scientific Reports 10 (2020).

[2] E. C. J. Tandoc, The facts of fake news: A research review, Sociology Compass 13 (2019) e12724.

[3] M. Mazza, G. Cola, M. Tesconi, Ready-to-(ab)use: From fake account trafficking to coordinated inauthentic behavior on twitter, Online Social Networks and Media 31 (2022) 100224.

[4] S. Tardelli, M. Avvenuti, M. Tesconi, S. Cresci, Characterizing social bots spreading financial disinformation, in: G. Meiselwitz (Ed.), Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis, Springer International Publishing, Cham, 2020, pp. 376–392.

[5] W. Marcellino, Detecting Conspiracy Theories on Social Media: Improving Machine Learning to Detect and Understand Online Conspiracy Theories, Technical Report, RAND Corporation, Santa Monica, CA, 2021.

[6] M. Gambini, S. Tardelli, M. Tesconi, The anatomy of conspiracy theorists: Unveiling traits using a comprehensive twitter dataset, Computer Communications 217 (2024) 25–40.

[7] S. Cresci, M. Petrocchi, A. Spognardi, M. Tesconi, R. D. Pietro, A criticism to society (as seen by twitter analytics), in: 2014 IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW), 2014, pp. 194–200.

[8] K. Shu, S. Wang, H. Liu, Understanding user profiles on social media for fake news detection, in: Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval, IEEE, Miami, FL, 2018, pp. 430–435.

[9] A. Giachanou, E. A. Ríssola, B. Ghanem, altri, The role of personality and linguistic patterns in discriminating between fake news spreaders and fact checkers, in: E. Métais, F. Meziane, H. Horacek, altri (Eds.), Natural Language Processing and Information Systems, Springer, New York, 2020, pp. 181–192.

[10] P. Zola, G. Cola, A. Martella, M. Tesconi, Italian top actors during the COVID-19 infodemic on Twitter, International Journal of Web Based Communities 18 (2022) 150–172.

[11] D. Mahl, M. S. Schäfer, J. Zeng, Conspiracy theories in online environments: An interdisciplinary literature review and agenda for future research, New Media & Society 25 (2023) 1781–1801.
[12] T. Mitra, S. Counts, J. W. Pennebaker, Understanding anti-vaccination attitudes in social media, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 10, 2016, pp. 269–278.
[13] S. A. Memon, K. M. Carley, Characterizing covid-19 misinformation communities using a novel twitter dataset, arXiv preprint arXiv:2008.00791 (2020).
[14] M. Schmitz, G. Muri´c, K. Burghardt, A python package to detect antivaccine users on twitter, arXiv preprint arXiv:2110.11333 (2021).
[15] A. G. Jiménez, Ángel Panizo-LLedot, J. Torregrosa, D. Camacho, Representational learning for the detection of covid-related conspiracy spreaders in online platforms, in: MediaEval'22: Multimedia Evaluation Workshop, CEUR Workshop Proceedings, CEUR-WS.org, Bergen, Norway and Online, 2023.
[16] H. Batzdorfer, H. Steinmetz, M. Biella, M. Alizadeh, Conspiracy theories on twitter: Emerging motifs and temporal dynamics during the covid-19 pandemic, International Journal of Data Science and Analytics 13 (2022) 315–333.
[17] J. Zeng, M. S. Schäfer, Conceptualizing "dark platforms": Covid-19-related conspiracy theories on 8kun and gab, Digital Journalism 9 (2021) 1321–1343.
[18] C, P. Clutton, A. G. Dunn, Pathways to conspiracy: The social and linguistic precursors of involvement in reddit's conspiracy theory forum, PLoS ONE 14 (2019).
[19] A. Fong, J. Roozenbeek, D. Goldwert, S. Rathje, S. van der Linden, The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on twitter, Group Processes & Intergroup Relations 24 (2021) 606–623.
[20] A. Bessi, Personality traits and echo chambers on facebook, Computers in Human Behavior 65 (2016) 319–324.
[21] A. Giachanou, B. Ghanem, P. Rosso, Detection of conspiracy propagators using psycho-linguistic characteristics, Journal of Information Science 49 (2023) 3–17.
[22] M. Donaldson, Plutchik's wheel of emotions, SixSeconds, 2022. URL: https://www.6seconds.org/2022/03/13/plutchik-wheel-emotions/, accessed: 2024-04-20.
[23] S. G. Tesfagergish, J. Kapočiūtė-Dzikienė, R. Damaševičius, Zero-shot emotion detection for semi-supervised sentiment analysis using sentence transformers and ensemble learning, Applied Sciences 12 (2022).