Liars know how to argue: An approach to disinformation analysis based on argument mining

Jose A. Diaz-Garcia[1][0000−0002−9263−1402], M. Dolores Ruiz[1][0000−0003−1077−3173], and Maria J. Martin-Bautista[1][0000−0002−6973−477X]

Department of Computer Science and Artificial Intelligence, University of Granada,
Daniel Saucedo Aranda, s/n, 18014 Granada
{jagarcia, mdruiz, mbautis}@decsai.ugr.es

Abstract. The ongoing digital era is significantly impacted by the dissemination of fake news and disinformation. Addressing this issue involves delving into the realm of Artificial Intelligence, where Natural Language Processing (NLP) stands out as one of the most active areas capable of contributing to solutions. In this paper, we present an analysis grounded in argument mining, with a specific focus on understanding the creation of fake pieces of content in social media. Our research reveals that deceptive narratives in social media often incorporate a substantial component of arguments. This finding not only sheds light on the intricacies of misinformation but also provides valuable insights for future research in combating this pervasive issue.

Keywords: argument mining · disinformation · social media analysis · NLP

1 Introduction

In the current digital era, the vast majority of the global population regularly utilizes social networks to gather information on various topics of interest. News and online social networks serve as essential platforms for numerous companies and governments, facilitating the dissemination of their ideas, products, and information. Many individuals leverage these channels to form their own opinions and ideas. However, the widespread success of these platforms has led to the emergence of fake accounts disseminating disinformation, with the aim of influencing public opinion.

The automated and orchestrated dissemination of deceptive content poses a significant challenge that numerous companies and countries are suffering. This issue becomes particularly pronounced in critical processes such as elections, where it can be manipulated to orchestrate massive campaigns of fake news [5, 17, 23]. Confronting this problem represents a formidable challenge, requiring the collaboration of diverse fields, including journalism and Artificial Intelligence (AI). Notably, Artificial Intelligence stands out as the primary means to detect deceptive content due to its inherent capacity to address the rapid generation and dissemination of deceptive content.

The continuous advancement of AI technologies provides a proactive defense against the evolving tactics employed by those trying to manipulate public opinion through disinformation campaigns. Within AI, NLP techniques contribute significantly to the detection and analysis of disinformation. Named Entity Recognition (NER)[9, 2], sentiment analysis [3], and language modeling [1] are among the key NLP methodologies that enhance the accuracy and depth of disinformation detection systems. Leveraging these techniques led us to a better understanding of textual content, enabling more effective identification of deceptive narratives. In the realm of NLP, argument mining emerges as a pivotal technique in our approach to disinformation detection and profiling. By applying argument mining to social network conversations, we aim to uncover the strategic use of argumentation in disseminating misinformation.

Argument mining [15] is an NLP technique that involves the extraction and analysis of arguments from textual data. The primary goal of argument mining is to identify and understand the structure of arguments within a given text, be it written articles, social media posts, or other forms of communication. The process typically involves the identification of premises, claims, and relationships between different components of an argument. To perform argument mining, various computational methods are employed, leveraging techniques such as NLP, machine learning, and linguistic analysis. Sentences are parsed to identify key elements, and relationships between these elements are established to construct the argumentative structure. This process enables the identification of persuasive language, opinion expressions, and reasoning patterns within the text.

In this paper, we make significant contributions to the field of disinformation detection and profiling. We introduce an innovative approach that harnesses argument mining to categorize and profile how disseminators of misinformation employ argumentation in social network conversations. The primary objective of our paper is to delineate differences in the length and presence of argumentation structures within fake and real content disseminated through social networks. Aligned with this objective, our key contributions include:

- A comprehensive analysis over the impact of argumentative elements on deceptive discourse within social media conversations.
- A comparative exploration of the prevalence of arguments between long and short texts.

The paper is organized as follows: Next section focuses on the study of related work. In Section 3 we go into detail into the argument mining proposal. In Section 4, we provide the results of the experimentation. Finally, in Section 5 we examine the conclusions and the future work.

2 Related works

While the field of argument mining in social networks has seen considerable research in several topics [12, 6, 16], there remains ample opportunity for further enhancements and applications. In their work [19], the authors offer a comprehensive review of various applications and approaches employed by argument mining within the context of Twitter, analysing the evolving landscape of this research domain. In the specific domain of argument mining applied to political and social sciences, a noteworthy contribution is made by Vecchi et al. in [20].

In [13], Habernal and Gurevych presented a comprehensive approach to argument mining, addressing challenges associated with processing web discourses characterized by noisy user-generated content. Their contribution included the introduction of a valuable collection of gold examples and the proposition of a methodology for argument mining, leveraging the power of machine learning techniques. Additionally, Visser et al. [21] presented a noteworthy dataset, US2016, designed for the analysis of argumentation in online discourses related to the 2016 US presidential elections. This dataset, annotated using the OVA Software [18], focuses on argumentative relations within television debates.

Following that line of argument mining in online discourses in [11], authors present a model aimed at enhancing opinion mining on Twitter. Their approach involves analyzing argument pieces through the construction of opinion trees using argumentative structures. The authors highlight the influential role of argumentation, emphasizing its utility in various NLP applications. For instance, they showcase its effectiveness in opinion mining, while our work extends its application to the domains of disinformation profiling and analysis.

Furthermore, in [10] authors introduced two novel challenges in the realm of argument mining within social networks, particularly on platforms like Twitter. These challenges involve the recognition of facts and the identification of sources. The authors approached these tasks as classification problems, employing the DART dataset [4] (an annotated collection of tweets designed for argument mining). Their findings suggest that this approach and challenges can be extended to other issues, such as fact-checking.

In [14] authors propose a fake news detection system in which they use subjective elements of language including argumentation to create a fake news classifier. This paper is closely related to our proposal, although it explores not only argumentation but also other factors such as feelings or presuppositions.

Building upon prior research, it is evident that argument mining serves as a valuable tool for profiling online discourses. In our study, we model and profile the impact of these argumentative structures within both fake and real or fact-checked tweets. To the best of our knowledge, our research represents the first approach that addresses disinformation profiling through the application of argument mining techniques.

3 Methodology

In this section, we elucidate our methodology for mining and analyzing arguments within both real and fake tweets.

3.1 Data

For our analysis, we have concentrated on the TruthSeeker dataset [8]. This recently proposed dataset stands out as the most comprehensive collection in the domain of profiling fake and real content in social networks to date. Comprising 134,198 tweets sourced from news categorized by PolitiFact as either real or false. Experts start by extracting keywords for each news item and subsequently crawl tweets associated with each keyword. Using a crowdsourcing method of voting, a dataset of fake and real tweets based on the news is meticulously curated for each tweet.

Our focus lies in analyzing the arguments within the dataset. To conduct a more thorough examination, considering that larger texts offer increased potential for argumentative content, we have bifurcated the dataset twice. The initial division is based on the distribution of word counts, categorizing texts with 40 or more words as 'large texts' and the remaining as 'short texts' (Figure 1). The rationale behind this decision is to maintain a balanced distribution of examples for each split. For each subset, aiming to discern differences in argumentation between deceptive and non-deceptive content, we further partitioned the dataset into true and false content. For each partition of the dataset, we conducted argument mining to discern differences in terms of argument presence and, if an argument is detected, to analyze its length.

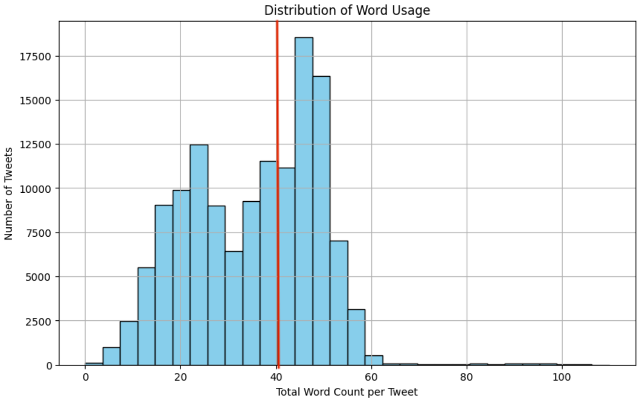


Fig. 1: Word count distribution and selection for experimentation
sch

3.2 Argument mining

For the extraction and analysis of argumentative components in our texts, our focus lies in comparing different elements of an argument.

- **Arg-B** represents the initial word indicating the start of an argument.

- **Arg-I** denotes a word that is part of an argument.
- **O** indicates that a word is not part of an argument.

As an illustration, consider the tweet '*@6d6f636869 Not as many people are literally starving and out in the streets as they were in the 19th century. Isn't capitalism grand? Meanwhile, we're facing an eviction moratorium threatening to make millions of Americans homeless. Fuck off with this corporatist propaganda.*' In this instance, we can conduct argument mining to obtain its tokens as follows:

[('@6d6f636869', 'Arg-B'), ('Not', 'Arg-I'), ('as', 'Arg-I'), ('many', 'Arg-I'), ('people', 'Arg-I'), ('are', 'Arg-I'), ('literally', 'Arg-I'), ('starving', 'Arg-I'), ('and', 'Arg-I'), ('out', 'Arg-I'), ('in', 'Arg-I'), ('the', 'Arg-I'), ('streets', 'Arg-I'), ('as', 'Arg-I'), ('they', 'Arg-I'), ('were', 'Arg-I'), ('in', 'Arg-I'), ('the', 'Arg-I'), ('19th', 'Arg-I'), ('century', 'Arg-I'), ('.', 'Arg-I'), ('Isnt', 'Arg-I'), ('capitalism', 'Arg-I'), ('grand', 'Arg-I'), ('?', 'Arg-I'), ('Meanwhile', 'Arg-I'), (',', 'Arg-I'), ('were', 'Arg-I'), ('facing', 'Arg-I'), ('an', 'Arg-I'), ('eviction', 'Arg-I'), ('moratorium', 'Arg-I'), ('threatening', 'Arg-I'), ('to', 'Arg-I'), ('make', 'Arg-I'), ('millions', 'Arg-I'), ('of', 'Arg-I'), ('Americans', 'Arg-I'), ('homeless', 'Arg-I'), ('.', 'O'), ('Fuck', 'O'), ('off', 'O'), ('with', 'O'), ('this', 'O'), ('corporatist', 'O'), ('propaganda', 'O'), ('.', 'O')]

Through this simple analysis, we can gain a wealth of valuable insights. For instance, a higher frequency of **Arg-B** indicates a greater number of arguments and argumentative tweets. Similarly, an increased occurrence of **Arg-I** implies arguments composed of more words or more intricate structures, which usually implies better arguments. For argument mining, we employed Canary [7], a tool that harnesses various NLP functions and tools, amalgamating them to efficiently compute argumentative structures within a text in a straightforward manner. It is worth mentioning that before extracting argumentative structures, we performed traditional text preprocessing steps to ensure the cleanliness of the tweet, all the while retaining valuable information essential for argument mining.

4 Results

This section presents the outcomes of applying argument mining and segmentation techniques to various configurations of the TruthSeeker dataset based on the text length. Our initial analysis focuses on the classification of tweets containing the presence of **Arg-B** structures, indicating the existence of at least one argumentative structure. The results for short texts, consisting of less than 40 words, are detailed in Table 1, while Table 2 provides the results for large texts.

Upon closer examination of the results, an unexpected trend emerges: deceptive tweets exhibit a higher degree of argumentation compared to their real counterparts. This contrast is particularly pronounced in short texts, with 3.57% more fake tweets featuring argumentative structures than real short tweets. In longer texts, the difference is slightly reduced, standing at 2.83%, which can

Label	**Argument Prediction**	**Number**	**%**
Real	True	26984	68,286
Real	False	12532	31,713
	Total	39516	100
Fake	True	24420	71.853
Fake	False	9566	28.146
	Total	33986	100

Table 1: Volume of fake and real tweets with argumentative structures over short texts dataset

Label	**Argument Prediction**	**Number**	**%**
Real	True	20549	69.730
Real	False	8920	30.269
	Total	29469	100
Fake	True	22661	72.568
Fake	False	8566	27.431
	Total	31227	100

Table 2: Volume of fake and real tweets with argumentative structures over large texts dataset

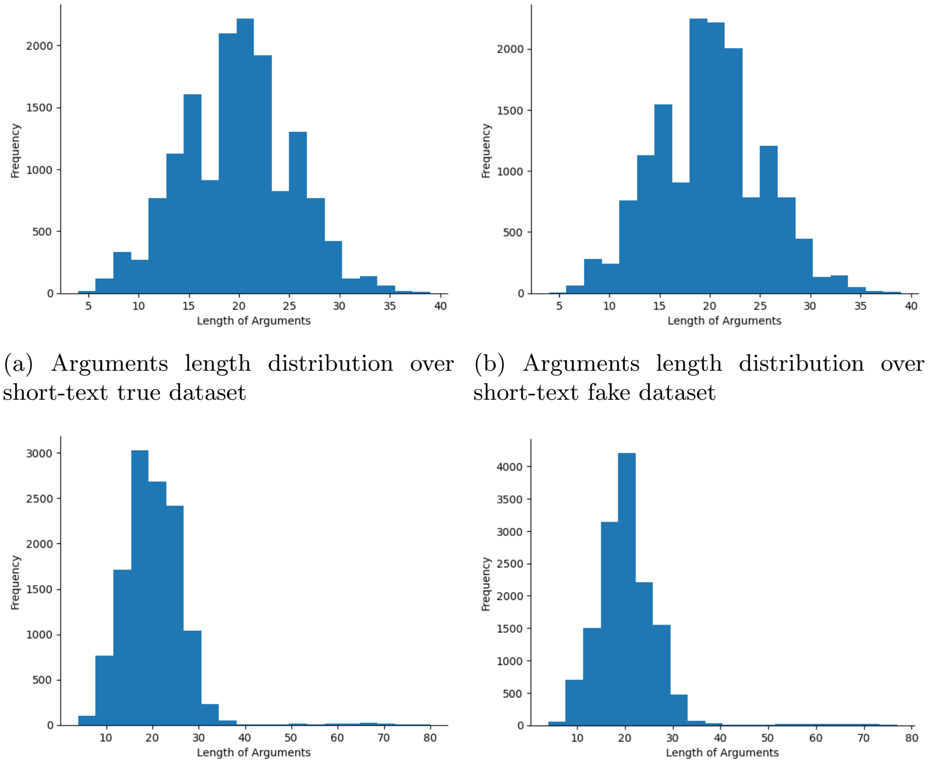
be attributed to the increased likelihood of engaging in argumentation within lengthier textual contexts.

After scrutinizing the presence of **Arg-B** across diverse datasets, our attention shifted towards evaluating the complexity of arguments. In our context, a complex argument is delineated by the number of words or elements within its argumentative structure. Specifically, an increase in **Arg-I** components implies a longer argument, and, in our interpretation, greater length signifies greater complexity. It is noteworthy to mention that, owing to the distinctive use of punctuation marks in user-generated text, we have excluded these marks from being considered in the analysis of argument segmentation and detection at this stage.

Figure 2 illustrates the distribution of different lengths for **Arg-I** in tweets containing arguments, categorized by the length of the tweets and their nature as either genuine or deceptive content. It is essential to note that for short texts, the highest situation entails only 40 words, enabling the creation of a more detailed graph based on the bins of the bar plot. Conversely, in large-text datasets, a broader range of arguments is observed, ranging from 1 to 80. This is because we have exclusively coded tweets over 40 words, with 80 representing the maximum **Arg-I** sequence found. However, it is crucial to acknowledge that in this split dataset, a tweet categorized as long-text (e.g., 75 words) may have fewer than 40 **Arg-I** sequences, and that is the reason between the difference of graphs in terms of bins in the bar plot.

Upon scrutinizing the graphs, a consistent pattern emerges across all cases, with the mode consistently at 20, representing the most prevalent value for the

50th percentile. Notably, when examining datasets associated with authentic texts or tweets, an intriguing observation pertains to the **Arg-I** sequence length values. Here, we observe significantly elevated values extending from 20 to 30, indicating a noteworthy level of variability in argumentation within truthful content. This implies that content representing the truth tends to exhibit more diverse argumentative structures compared to deceptive content, which experiences a rapid decline after reaching its mode.



(a) Arguments length distribution over short-text true dataset

(b) Arguments length distribution over short-text fake dataset

(c) Arguments length distribution over large-text true dataset

(d) Arguments length distribution over large-text fake dataset

Fig. 2: Arguments length distribution over the different datasets

Another interesting analysis can be unraveled if we analyze the volume of argumentation from the point of view of the average length of **Arg-I** sequences. In Table 3 the averages are shown as a function of the configuration of each dataset. We can see, as in the two cases related to fake or misleading content, the averages are higher at a significance level e 0.05. This result, as did the analysis based on the appearance of **Arg-B**, again leads us to conclude that misleading content is composed of more arguments and more complex, which is related to the need to persuade [22] through arguments that are behind the objective of those who seek to influence public opinion through the dissemination of false content.

Dataset Configuration	Average Length
Fake Large Texts	20,43
Real Large Texts	20,25
Fake Short Texts	20,37
Real Short Texts	20,08

Table 3: Average length of argument structures (**Arg-I**)

5 Conclusion

In this paper, we conducted a comparative analysis to explore the distinctions between deceptive and real content in terms of argument presence, employing argument mining techniques. Contrary to expectations, our results revealed that deceptive content tends to exhibit a higher degree of argumentation compared to real content. In percentage terms, our analysis indicates a 3% higher prevalence of arguments in fake datasets compared to real datasets. Furthermore, we observed that arguments in deceptive discourses tend to be more intricate than those found in truthful discourses. Our in-depth investigation leads us to believe that this phenomenon is linked to the deceptive necessity of persuading individuals about their deceptive ideas. This work lays the groundwork for subsequent research in argument mining over fake and real content. One prospective avenue for our future work involves the imperative to introduce a more sophisticated layer for analyzing argument complexity. This entails considering various types of arguments, including divergent arguments, convergent arguments, or linked arguments. Additionally, another promising direction for future exploration is the development of new features rooted in argumentative presence or weighting. These features could be seamlessly integrated into the construction of a classifier, potentially enhancing the model's robustness in discerning between deceptive and authentic content and offering valuable insights into the evolving landscape of online discourse analysis.

Acknowledgment

The research reported in this paper was supported by the DesinfoScan project: Grant TED2021-129402B-C21 funded by MCIN/AEI/10.13039/501100011033 and, by the European Union NextGenerationEU/PRTR, and FederaMed project:

Grant PID2021-123960OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe. Finally, the research reported in this paper is also funded by the European Union (BAG-INTEL project, grant agreement no. 101121309).

References

1. Aggarwal, A., Chauhan, A., Kumar, D., Verma, S., Mittal, M.: Classification of fake news by fine-tuning deep bidirectional transformers based language model. EAI Endorsed Transactions on Scalable Information Systems **7**(27), e10–e10 (2020)
2. Al-Ash, H.S., Wibowo, W.C.: Fake news identification characteristics using named entity recognition and phrase detection. In: 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE). pp. 12–17. IEEE (2018)
3. Alonso, M.A., Vilares, D., Gómez-Rodríguez, C., Vilares, J.: Sentiment analysis for fake news detection. Electronics **10**(11), 1348 (2021)
4. Bosc, T., Cabrio, E., Villata, S.: DART: a dataset of arguments and their relations on Twitter. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 1258–1263. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://aclanthology.org/L16-1200
5. Bovet, A., Makse, H.A.: Influence of fake news in twitter during the 2016 us presidential election. Nature communications **10**(1), 7 (2019)
6. Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K., Hwang, A.: Ampersand: Argument mining for persuasive online discussions. arXiv preprint arXiv:2004.14677 (2020)
7. Christopher Wales, Calvin-Castle Gill, S.W.: Canary: A tool for argument mining. https://github.com/Open-Argumentation/Canary (Year Accessed)
8. Dadkhah, S., Zhang, X., Weismann, A.G., Firouzi, A., Ghorbani, A.A.: The largest social media ground-truth dataset for real/fake content: Truthseeker. IEEE Transactions on Computational Social Systems (2023)
9. De Magistris, G., Russo, S., Roma, P., Starczewski, J.T., Napoli, C.: An explainable fake news detector based on named entity recognition and stance classification applied to covid-19. Information **13**(3), 137 (2022)
10. Dusmanu, M., Cabrio, E., Villata, S.: Argument mining on twitter: Arguments, facts and sources. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2317–2322 (2017)
11. Grosse, K., Chesñevar, C.I., Maguitman, A.G.: An argument-based approach to mining opinions from twitter. AT **918**, 408–422 (2012)
12. Habernal, I., Faber, D., Recchia, N., Bretthauer, S., Gurevych, I., Spiecker genannt Döhmann, I., Burchard, C.: Mining legal arguments in court decisions. Artificial Intelligence and Law pp. 1–38 (2023)
13. Habernal, I., Gurevych, I.: Argumentation mining in user-generated web discourse. Computational Linguistics **43**(1), 125–179 (2017)
14. Jeronimo, C.L.M., Marinho, L.B., Campelo, C.E., Veloso, A., da Costa Melo, A.S.: Fake news classification based on subjective language. In: Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services. pp. 15–24 (2019)

15. Lawrence, J., Reed, C.: Argument mining: A survey. Computational Linguistics **45**(4), 765–818 (2020)
16. Mayer, T., Cabrio, E., Villata, S.: Transformer-based argument mining for healthcare applications. In: ECAI 2020, pp. 2108–2115. IOS Press (2020)
17. Ncube, L.: Digital media, fake news and pro-movement for democratic change (mdc) alliance cyber-propaganda during the 2018 zimbabwe election. African Journalism Studies **40**(4), 44–61 (2019)
18. REED, M., Janier, M., Lawrence, J.: Ova+: An argument analysis interface. In: Computational Models of Argument: Proceedings of COMMA. vol. 266, p. 463 (2014)
19. Schaefer, R., Stede, M.: Argument mining on twitter: A survey. it-Information Technology **63**(1), 45–58 (2021)
20. Vecchi, E.M., Falk, N., Jundi, I., Lapesa, G.: Towards argument mining for social good: A survey. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1338–1352 (2021)
21. Visser, J., Konat, B., Duthie, R., Koszowy, M., Budzynska, K., Reed, C.: Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. Language Resources and Evaluation **54**(1), 123–154 (2020)
22. Walton, D.: Deceptive arguments containing persuasive language and persuasive definitions. Argumentation **19**(2), 159–186 (2005)
23. Wang, T.L.: Does fake news matter to election outcomes?: The case study of taiwan's 2018 local elections. Asian Journal for Public Opinion Research **8**(2), 67–104 (2020)