# Leveraging Temporal Analysis to Predict the Impact of Political Messages on Social Media in Spanish

Ibai Guillén-Pacho

*Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain*

Abstract

Social networks such as TikTok, Facebook, or X introduce techniques to inform users if the content they are consuming may be fake. This, together with the account banning for hate speech or disinformation spread, is leading more and more pseudo-media in Spain to use Telegram to communicate with their audience. Thus, it is difficult to warn users about the veracity of the content, leading them to accept political disinformation as true if it aligns with their beliefs, which ultimately promotes their polarization. In this work, we want to *identify the political messages that will have the greatest impact on people* to recommend when it is necessary to initiate a refutation strategy if it is disinformative, so that refutation begins before disinformation is taken as true by a part of society. To estimate the impact of political messages, we take into account the polarization generated by them and their virality. Our main hypothesis is that this value is proportional to the time of publication, with the greatest impact in the most politically and socially sensitive contexts. Hence, our goal is to compile a dataset of political messages disseminated on Telegram (along with the generated responses) and its temporal context, in order to develop methods and metrics to identify the expected impact and when they might have the greatest impact.

**Keywords**

disinformation, fake news impact, polarization, response generation

## 1. Introduction

Disinformation is defined as “false information that is shared to intentionally mislead" and is one of the main information disorders [1]. The different information disorders, shown in Figure 1, are distinguished by the information they use (real or fake) and the intention with which the news are generated (harm, mislead, other). The main problem with misleading fake news, the ones that provoke disinformation, is that not only can they be of different types (hoax, rumor, conspiracy theory, etc.), but they can spread very quickly and have a huge impact. This can lead to major problems, in terms of health (e.g., suggesting methods of curing diseases ), as a polarizer of society (e.g., creating hoaxes about the Spanish law on transgender rights ), and as a way to promote hatred (e.g., against immigrants ), among others.

Disinformation presents a risk to political speech and democracy, the impact it has on citizens is mainly due to the way it goes viral and the polarization it induces [2]. The more viral and polarizing a news story is, the greater its impact. To mitigate this risk, [2] suggest that

*Doctoral Symposium on Natural Language Processing, 26 September 2024, Valladolid, Spain.*

ibai.guillen@upm.es (I. Guillén-Pacho)
https://iguillenp.github.io/iguillenp/ (I. Guillén-Pacho)
0000-0001-7801-8815 (I. Guillén-Pacho)

CEUR Workshop Proceedings (CEUR-WS.org)

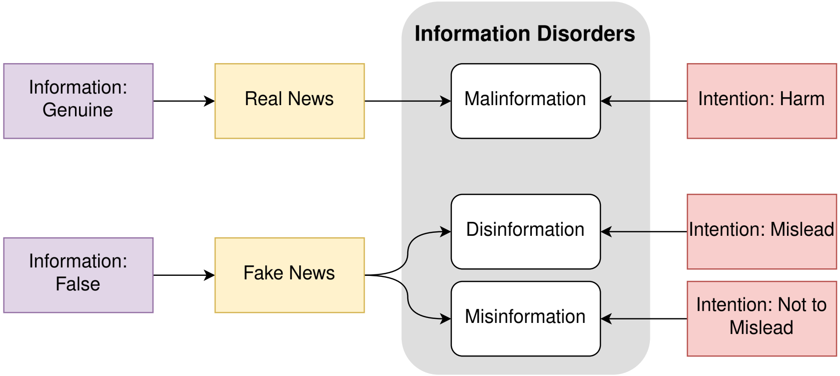

**Figure 1:** Information disorders based on [1].

the refutation should focus on high impact disinformation while omitting low impact cases. In this paper, we present an approach to predict the impact of political messages on social networks. The goal is to 1) help human annotators focus their resources on the most harmful messages and 2) help refutation mechanisms before the message gets a big impact on social media and biases public opinion. Furthermore, this method seeks to prevent the Streisand effect by unintentionally highlighting messages that would otherwise remain insignificant.

The structure of the paper is as follows. Section 2 discusses the background and related work. Section 3 details the proposed research, including assumptions, the research problem, hypothesis, research questions, and the objectives. Section 4 explains the research methodology adopted. Lastly, Section 5 summarizes the work done, the conclusions drawn, and suggests directions for future research.

## 2. Background and Related Work

Due to its low moderation and group/community-centered structure, Telegram provides an environment more prone to sharing political disinformation, which in turn implies a risk of political radicalization of users [3]. In Spain, Telegram is one of the social networks used by most of the main sources of disinformation. These pseudo-media, listed in [4], have their own public Telegram groups that they use to share disinformation[1]. In fact, some of these pseudo-media had accounts on other social networks that have been closed for spreading false information and promoting hatred towards minorities. They claim that these acts are censorship and encourage their readers to join their Telegram groups, making this social network one of the main channels for the dissemination of disinformation.

To combat this disinformation, we can differentiate three refutation strategies according to [5]: prebunking, debunking, and narrative counter. **Prebunking** focuses on protecting individuals from disinformation and attempts to influence them. The main techniques focus on analyzing the message and detecting its language intensity, identifying the source credibility,

[1] Have a Telegram group of their own: *Contando Estrelas*, *Mediterráneo Digital*, *Mpr21*, *El Diestro*, *Alerta Nacional*, and *Euskalnews*. No known Telegram group of their own: *El correo de España* and *Altavoz de Successos*

improving individuals' ability to deal with false information, etc. [6]. **Debunking** uses facts and reliable information to refute disinformation. Fact-checking for debunk messages can be done manually (with experts or crowd-sourcing) or with automated tools trained on reliable knowledge bases [7]. **Narrative counter** generates an easy to understand explanation without the need for empirical evidence. This strategy is also known as counterspeech and, although it can be done manually, its complexity makes it difficult to scale; thus, generative models represent a potential improvement [8]. Moreover, it is one of the most effective methods in reducing the effect of disinformation on the beliefs, intentions, and attitudes of individuals [5].

However, there are too many communities and too many messages to fight them all. Choosing the **highest impact messages** can maximize the effect of these refutation strategies without being resource intensive and without giving importance to irrelevant messages. As the impact of disinformation directly depends on its virality and the polarization it produces [2], it is important to predict these aspects to make the best decision.

On the one hand, **virality** in social networks is a phenomenon that has been widely studied in the literature. Studies have been conducted on its *propagation structure*, *measurement* and *prediction*. First, there are two main types of *propagation structures* according to [9]: diffusion, when virality increases as more people adopt and share the message; and broadcast, when virality is gained by sharing a single message that reaches many individuals. Second, the main approaches for *measuring* virality use engagement metrics: number of reposts, number of likes, number of followers of the author, combination of several, etc. [10]. Finally, for automatic *prediction* of virality, the content of the message and its characteristics (whether it contains media, whether the author is verified, has positive or negative sentiment, etc.) are usually used [10].

On the other hand, we refer as **generated polarization** to the polarization generated in the responses to the message and not in the content of the message itself. Therefore, to know this feature, it is necessary to simulate the discussion generated by the message, to later measure the polarization of the responses. This approach ties into the broader topic of dialogue systems (DS), where the generation of responses to a message is widely discussed. These systems are usually divided according to their main specific features [11]: *approach* (architecture of the system), *purpose* (general type of system) and *model* (underlying technique used).

The first feature (*approach*) depends on how the different components of a DS interact. The components form the architecture of a DS, and according to [12, 13], there are four main ones: natural language understanding (to identify user intents and key information), tracking of the dialogue (to monitor user belief states based on dialogue history), learning the dialogue policy (to decide the next action), and natural language generation (to produce the response of the system in the dialogue). If they are trained separately and interact with each other, the DS follows a *pipeline approach* (also referred to as *modular approach*). However, it is possible to use a single trained model playing the role of all components, in this case the DS follows an *end-to-end approach*. The second feature (*purpose*), depends on the capability of the DS to work in certain domains. If the DS is designed to be flexible and not limited to a particular task, it is *open-domain*. In contrast, if it is specialized in solving a specific task, it is *task-oriented*.

The last feature (*model*), constitutes the main technique used to build the DS. There are many different techniques, but for the latter purpose (*task-oriented*), which is in line with our work, the most promising ones are those based on *reinforced learning* (RL) [11] and *large language* *models* (LLM) [12]. We consider LLMs to be the best option for this domain because of their ability to understand and generate natural language and their ability to adapt to a context with few examples [12]. Nevertheless, it is necessary to fine-tune them with domain-specific information [12, 13] and combine them with RL [13] and quality prompts [11] to improve the quality of the results.

In the literature, applications of LLM to **generate responses** are usually related to detecting fake news because comments provide an important social context for the task [14, 15]. The main papers in this field provide LLMs with the content of the news item and additional information to generate comments, this additional information can be: an example comment [14], a description of a user profile to be emulated by the model [15], both [16], etc. However, in an approach that has to work with continuous data streams to detect the impact of political disinformation, extracting the content of all the news items to be analyzed is an inacceptable resource-intensive process. To optimize resources, we believe the best approach for this type of information is to use only the news headline, which limits the context that can be given to the model and makes the problem analogous to generating responses to social media messages. Within this other domain, we find several resources (fine-tuned LLMs [17], datasets [18], frameworks [19], etc.) and different methods for providing context to the LLM in generating responses, such as: specifying which sentiment to focus on [20]; evidence on which to base the response [21]; intention or emotion of the speaker or of the response [19]; etc. In short, all these approaches try to generate responses that are as real as possible, based on a piece of text and a context.

The realism of the responses generated is crucial for **measuring polarization**, as this phenomenon occurs when individuals form opinion groups that interact little or no with each other (also known as cyberbalkanization) [22]. It is common to find works that measure polarization manually, both at the user level, by surveying users and asking them about their position on a predefined set of issues (e.g. [23]); and at text level, analyzing the content following steps defined in a framework (e.g. [24]). Increasing efforts are being made to automate this measurement. For example, at the user level, there is a method that assesses the polarization of users and groups by analyzing interactions between users, the content of their discussions, and the opinion leaders or sources to which they refer [25]. We also found a study that estimates content polarization through sentiment analysis. The general sentiment of all content is calculated and the sentiment of each content is taken individually; the greater the difference between these two values, the greater the polarization of the content [26].

In short, to help fight against political disinformative messages and prevent further polarization of society, sophisticated tools are needed. Tools capable of managing the different elements of the political discourse to prevent radicalization. For this, it is essential to know the impact of political messages on people, a task that requires measuring the virality of the messages and the polarization of individuals reactions to them. In this way, disinformative political messages can be refuted before having a major impact, avoiding the bias that exposure to false information creates in people's beliefs.

## 3. Proposed Research

Based on the previous analysis of the background and related work, we **assume** that political disinformation carries multiple risks to society, such as political radicalization and normalization of hate speech toward minorities. To mitigate its effects, there are different manual and automatic refutation strategies, but, due to the large volume of messages created every day, analyzing and refuting all of them is not viable. In this work, we want to solve the **research problem** of *identifying political messages that will have a great impact on people.* For this, our **research hypothesis** is that the *impact of a political message is greater when it is published during times of high social or political sensitivity*, such as elections, social movements, national crises, etc. The different **research questions** (RQ) that arise as a result of the above are:

RQ1: How does the virality of a political message vary according to the temporal context in which it is published?

RQ2: How can LLMs emulate and predict the dynamics of the social media debate provoked by a political message?

RQ3: How can the social impact of political message be measured at different points in time?

RQ4: What are the periods of high social or political sensitivity during which the publication of a political message has the greatest impact?

To address the above RQs, we define several **objectives** (O) and **sub-objectives** (SO) to be satisfied. Although the first objective is to generate the necessary resources and is a requirement, the rest are related to the following RQs, O2 to RQ1, O3 to RQ2, and O4 to RQ3 and RQ4. The list of objectives is as follows:

O1: Generate the resources needed.

- SO1-1: Build a corpus of disinformative political messages in Spanish spread on Telegram.
- SO1-2: Represent the social and political context when each message was published.

O2: Predict the virality of political messages.

- SO2-1: Train a baseline model to predict the virality of political message in the SO1-1 corpus relying only in the content of the message.
- SO2-2: Explore which elements of the context captured in SO1-2 improve the performance of the model.
- SO2-3: Build a model to predict virality with the best configuration found.

O3: Emulate the debate that could be provoked by political messages.

- SO3-1: Evaluate the performance of LLMs in predicting existing responses to the resource built in O1S1.
- SO3-2: Explore which elements of the context captured in SO1-2 improve the performance of the best LLM analysed.
- SO3-3: Build a DS that injects the selected context elements in SO3-2 together with the political message into the LLM to generate the responses.

O4: Create a method to automatically estimate whether or not a political message will have a great impact and when it is expected to have the greatest impact.

SO4-1: Estimate the polarisation of the emulated conversation in O3 for each message item.

SO4-2: Calculate the impact of the message with the virality calculated in O2 and the polarity it generates in the responses produced in O3.

SO4-3: Divide the messages into topics and analyze whether their impact follows a temporal or social pattern.

SO4-4: Group the above steps together to generate a system that estimates the impact of political messages at given moments.

## 4. Methodology

We organize the research in four main phases: 1) analysis and study of the domain and existing solutions, 2) defining the problem, the research and its objectives, and 3) experimentation and evaluation. In the first phase, we **study the domain** of knowledge representation (for O1), the evaluation of virality in social networks (for O2), the generation of responses to posts and news (for O3), and the analysis of the temporal evolution of knowledge (for O4). In addition, we also analyze the main **existing solutions**, such as Knowledge Graphs for knowledge representation, engagement metrics to measure virality, LLM to generate responses to comments, and concept drift to analyze the evolution of information. In the second phase, we **define the problem** and the research line that we will follow to solve it, establishing the hypotheses, the research questions and the objectives. Finally, in the third phase, we design the **experiments** to be carried out and their evaluation to meet the objectives and answer the research questions. The organization of work to meet the objectives is shown in Figure 2.

## 5. Conclusions and Future Work

The main work done is focused on generating the necessary resources (O1). The main difficulty of this objective is to represent the social and political context of each news item. Our approach is to generate an ontology (to be published) and build a Knowledge Graph so that the social and political context at a specific moment in time can be obtained through queries. We are also working on predicting the virality of political messages on social networks. We conduct experiments to analyze which elements of the context are the most useful to improve the accuracy of the prediction (ideology of the speaker, presence of hate speech, if it is tagged as disinformative, whether it is published during the election campaign, etc.).

Context has proven to be a key element in our preliminary analysis of virality prediction and we believe it will also be a key element in the generation of responses. As mentioned above, responses to political messages are essential to understand the impact they have on society, so we believe that this work can improve existing systems in the fight against political disinformation. Although the research may be ambitious, we hope to have a system capable of recommending when to refute a disinformative political message and when a message is expected to have the greatest impact.

We are now past the halfway point of the 2nd year and will continue in accordance with Figure 2. The next steps will be the prediction of virality and the generation of responses to news. In both objectives, the state of the art will be further explored, and the best approach to tackle the tasks will be decided. We will explore how context affects the models used and what information is most relevant to achieve quality results.

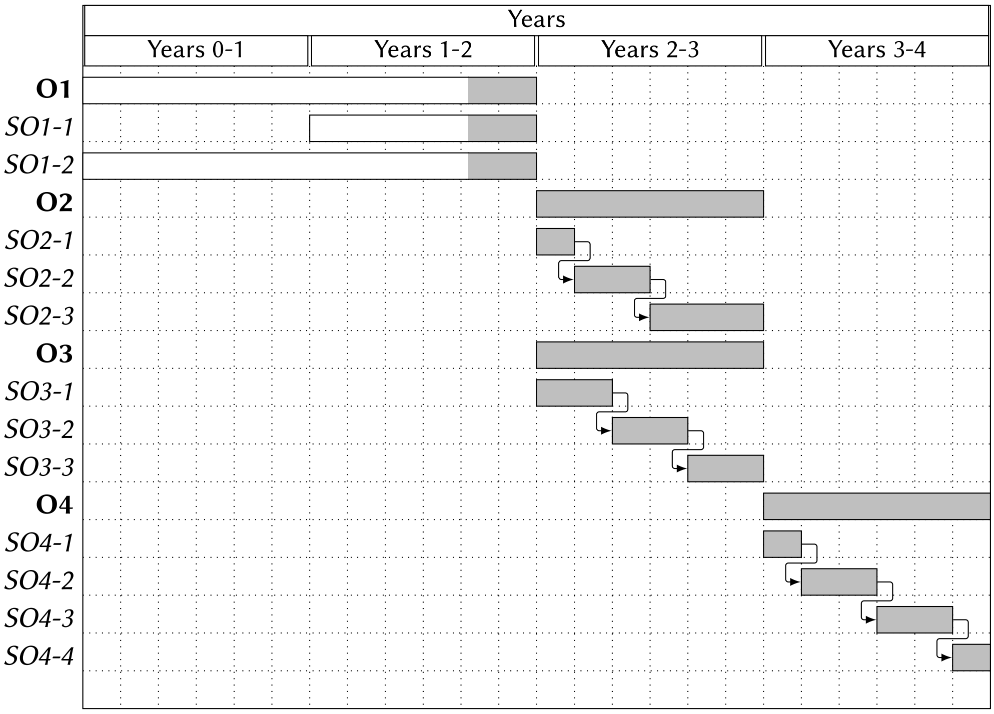

**Figure 2:** Timeline of the research.

## Acknowledgments

This work is supported by the Predoctoral Grant (PIPF-2022/COM-25947) of the Consejería de Educación, Ciencia y Universidades de la Comunidad de Madrid, Spain. First, I would like to thank my supervisors Carlos Badenes Olmedo and Óscar Corcho García. Second, I gratefully acknowledge the Universidad Politécnica de Madrid (www.upm.es) for providing computing resources on Magerit Supercomputer.

## References

[1] E. Aïmeur, S. Amri, G. Brassard, Fake news, disinformation and misinformation in social media: A review, Social Network Analysis and Mining 13 (2023) 30. doi:10.1007/s13278-023-01028-5.

[2] A. French, V. C. Storey, L. Wallace, A typology of disinformation intentionality and impact, Information Systems Journal n/a (2023). doi:10.1111/isj.12495.

[3] J. Rieskamp, M. Mirbabaie, M. Langer, A. Kocur, From Virality to Veracity: Examining False Information on Telegram vs. Twitter (2024). URL: https://hdl.handle.net/10125/106687.

[4] D. P. Sampio, A. Carratalá, Injecting disinformation into public space: pseudo-media and reality-altering narratives, Profesional de la información / Information Professional 31 (2022). doi:10.3145/epi.2022.may.12.

[5] M. A. Amazeen, A. Krishna, Refuting misinformation: Examining theoretical underpinnings of refutational interventions, Current Opinion in Psychology 56 (2024) 101774. doi:10.1016/j.copsyc.2023.101774.

[6] C. D. Boman, Examining characteristics of prebunking strategies to overcome PR disinformation attacks, Public Relations Review 47 (2021) 102105. doi:10.1016/j.pubrev.2021.102105.

[7] M. Soprano, K. Roitero, D. La Barbera, D. Ceolin, D. Spina, G. Demartini, S. Mizzaro, Cognitive Biases in Fact-Checking and Their Countermeasures: A Review, Information Processing & Management 61 (2024) 103672. doi:10.1016/j.ipm.2024.103672.

[8] Y.-L. Chung, G. Abercrombie, F. Enock, J. Bright, V. Rieser, Understanding Counterspeech for Online Harm Mitigation, 2023. doi:10.48550/arXiv.2307.04761.

[9] S. Goel, A. Anderson, J. Hofman, D. J. Watts, The Structural Virality of Online Diffusion, Management Science 62 (2016) 180–196. doi:10.1287/mnsc.2015.2158.

[10] T. Elmas, S. Stephane, C. Houssiaux, Measuring and Detecting Virality on Social Media: The Case of Twitter's Viral Tweets Topic, Companion Proceedings of the ACM Web Conference 2023 (2023) 314–317. doi:10.1145/3543873.3587373.

[11] A. Algherairy, M. Ahmed, A review of dialogue systems: Current trends and future directions, Neural Computing and Applications 36 (2024) 6325–6351. doi:10.1007/s00521-023-09322-1.

[12] L. Qin, W. Pan, Q. Chen, L. Liao, Z. Yu, Y. Zhang, W. Che, M. Li, End-to-end Task-oriented Dialogue: A Survey of Tasks, Methods, and Future Directions, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 5925–5941. doi:10.18653/v1/2023.emnlp-main.363.

[13] H. Wang, L. Wang, Y. Du, L. Chen, J. Zhou, Y. Wang, K.-F. Wong, A Survey of the Evolution of Language Model-Based Dialogue Systems (2023). doi:10.48550/ARXIV.2311.16789.

[14] Y. Yanagi, R. Orihara, Y. Sei, Y. Tahara, A. Ohsuga, Fake news detection with generated comments for news articles, in: 2020 IEEE 24th International Conference on Intelligent Engineering Systems (INES), 2020, pp. 85–90. doi:10.1109/INES49302.2020.9147195.

[15] Q. Nan, Q. Sheng, J. Cao, B. Hu, D. Wang, J. Li, Let silence speak: Enhancing fake news detection with generated comments from large language models, 2024. doi:10.48550/arXiv.2405.16631. arXiv:2405.16631.

[16] H. Wan, S. Feng, Z. Tan, H. Wang, Y. Tsvetkov, M. Luo, DELL: Generating Reactions and Explanations for LLM-Based Misinformation Detection, 2024. doi:10.48550/arXiv.2402.10426.

[17] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, B. Dolan, DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation, in:

A. Celikyilmaz, T.-H. Wen (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 270–278. doi:10.18653/v1/2020.acl-demos.30.

[18] P. Zhong, C. Zhang, H. Wang, Y. Liu, C. Miao, Towards persona-based empathetic conversational models, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6556–6566. doi:10.18653/v1/2020.emnlp-main.531.

[19] W. Li, Y. Yang, P. Tuerxun, X. Fan, Y. Diao, A Response Generation Framework Based on Empathy Factors, Common Sense, and Persona, IEEE Access 12 (2024) 26819–26829. doi:10.1109/ACCESS.2024.3365533.

[20] Z. Yang, Z. Ren, W. Yufeng, S. Peng, H. Sun, X. Zhu, X. Liao, Enhancing Empathetic Response Generation by Augmenting LLMs with Small-scale Empathetic Models (2024). doi:10.48550/ARXIV.2402.11801.

[21] Z. Yue, H. Zeng, Y. Lu, L. Shang, Y. Zhang, D. Wang, Evidence-Driven Retrieval Augmented Response Generation for Online Misinformation (2024). doi:10.48550/ARXIV.2403.14952.

[22] T. Jiang, Studying opinion polarization on social media, Social Work and Social Welfare 4 (2022) 232–241. doi:10.25082/SWSW.2022.02.003.

[23] J. Lee, Y. Choi, Effects of network heterogeneity on social media on opinion polarization among south koreans: Focusing on fear and political orientation, International Communication Gazette 82 (2020) 119–139. doi:10.1177/1748048518820499.

[24] N. D. Goet, Measuring polarization with text analysis: Evidence from the uk house of commons, 1811–2015, Political Analysis 27 (2019) 518–539. doi:10.1017/pan.2019.2.

[25] S. C. Phillips, J. Uyheng, K. M. Carley, A high-dimensional approach to measuring online polarization, Journal of Computational Social Science 6 (2023) 1147–1178. doi:10.1007/s42001-023-00227-6.

[26] I.-J. Serrano-Contreras, J. García-Marín, Ó. G. Luengo, Measuring Online Political Dialogue: Does Polarization Trigger More Deliberation?, Media and Communication 8 (2020) 63–72. doi:10.17645/mac.v8i4.3149.