Divergent Emotional Patterns in Disinformation on Social Media? An Analysis of Tweets and TikToks About the DANA in Valencia

Iván Arcos[1], Paolo Rosso[1,2] and Ramón Salaverría[3]
[1]*PRHLT Research Center, Universitat Politècnica de València, Valencia, Spain*
[2]*ValgrAI Valencian Graduate School and Research Network of Artificial Intelligence, Spain*
[3]*Sc ain*
iarcgab@etsinf.upv.es, prosso@dsic.upv.es, rsalaver@unav.es

Keywords: Disinformation, DANA, Extreme Weather Events, Social Media, Journalism, Computational Linguistics, Emotional Patterns.

Abstract: This study investigates the dissemination of disinformation on social media platforms during the DANA event (DANA is a Spanish acronym for *Depresión Aislada en Niveles Altos*, translating to high-altitude isolated depression) that resulted in extremely heavy rainfall and devastating floods in Valencia, Spain, on October 29, 2024. We created a novel dataset of 650 TikTok and X posts, which was manually annotated to differentiate between disinformation and trustworthy content. Additionally, a Few-Shot annotation approach with GPT-4o achieved substantial agreement (Cohen's kappa of 0.684) with manual labels. Emotion analysis revealed that disinformation on X is mainly associated with increased sadness and fear, while on TikTok, it correlates with higher levels of anger and disgust. Linguistic analysis using the LIWC dictionary showed that trustworthy content utilizes more articulate and factual language, whereas disinformation employs negations, perceptual words, and personal anecdotes to appear credible. Audio analysis of TikTok posts highlighted distinct patterns: trustworthy audios featured brighter tones and robotic or monotone narration, promoting clarity and credibility, while disinformation audios leveraged tonal variation, emotional depth, and manipulative musical elements to amplify engagement. In detection models, SVM+TF-IDF achieved the highest F1-Score, excelling with limited data. Incorporating audio features into *roberta-large-bne* improved both Accuracy and F1-Score, surpassing its text-only counterpart and SVM in Accuracy. GPT-4o Few-Shot also performed well, showcasing the potential of large language models for automated disinformation detection. These findings demonstrate the importance of leveraging both textual and audio features for improved disinformation detection on multimodal platforms like TikTok.

1 INTRODUCTION

In the digital era, social media platforms such as X (formerly Twitter) and TikTok revolutionized communication and information dissemination. These platforms facilitate instant sharing and broad interactions, enabling users to engage with content in real-time and reach vast audiences with unprecedented speed. While this democratization of information access offers numerous benefits, it also presents significant challenges, particularly concerning the spread of false or misleading information.

Disinformation, defined by European Commission (2019) as "verifiably false or misleading information created, presented, and disseminated for financial gain or intentional deception of the public", has become a pervasive issue on social media. Unlike misinformation, which involves the unintentional

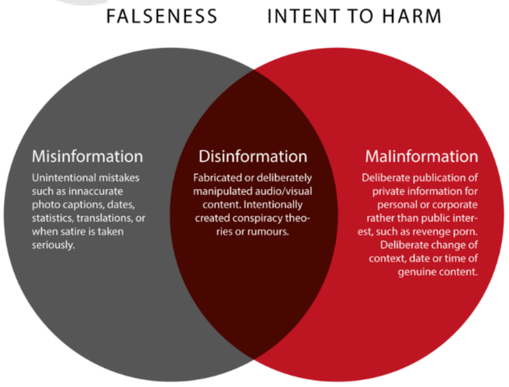

FALSENESS
INTENT TO HARM
Misinformation
Unintentional mistakes such as innaccurate photo captions, dates, statistics, translations, or when satire is taken seriously.
Disinformation
Fabricated or deliberately manipulated audio/visual content. Intentionally created conspiracy theories or rumours.
Malinformation
Deliberate publication of private information for personal or corporate rather than public interest, such as revenge porn. Deliberate change of context, date or time of genuine content.

Figure 1: Relationship among Disinformation, Misinformation, and Malinformation (Wardle and Derakhshan, 2017).[1]

[1] https://firstdraftnews.org/latest/coe_infodisorder/

Arcos, I., Rosso, P. and Salaverría, R.
Divergent Emotional Patterns in Disinformation on Social Media? An Analysis of Tweets and TikToks About the DANA in Valencia.
DOI: 10.5220/0013392800003890
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 1*, pages 925-936
ISBN: 978-989-758-737-5; ISSN: 2184-433X
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

sharing of incorrect information without malicious intent, disinformation is strategically crafted to manipulate public perception, influence opinions, and achieve specific agendas. Additionally, malinformation refers to the dissemination of genuine information with the intent to harm or manipulate, further complicating the information ecosystem. These three categories are illustrated in Figure 1.

The design of social media platforms is characterized by algorithms that prioritize engagement, the formation of echo chambers that reinforce existing beliefs, and the ease of content sharing. These features facilitate the rapid and widespread propagation of disinformation. This environment allows malicious actors, including those whose aim is to spread disinformation to increase their number of followers, as well as automated accounts and coordinated misinformation campaigns, to effectively amplify false narratives. Furthermore, the viral nature of social media content means that disinformation can quickly reach and influence large segments of the population, often outpacing fact-checking and regulatory efforts aimed at curbing its spread.

Understanding the mechanisms behind the dissemination of disinformation is crucial for developing effective countermeasures. Emotional and linguistic patterns play a significant role in how disinformation resonates with audiences and spreads across platforms. Analyzing these patterns can provide insights into the narrative employed by those who create and distribute false information, as well as the emotional triggers that make such content more likely to be shared and believed (McLoughlin et al., 2024).

This study focuses on the spread of disinformation during the DANA event in Valencia, Spain, on October 29, 2024, which caused severe flooding, resulting in 224 deaths and three missing persons. By examining emotional and linguistic patterns in 650 annotated posts from TikTok and X, we aim to elucidate the strategies used in disseminating false narratives and their impact on public sentiment. The research employs both traditional machine learning models and advanced language models, such as transformers and Large Language Models (LLMs), to detect disinformation, highlighting the strengths and limitations of each approach.

The remainder of this article is organized as follows: Section 2 reviews related literature on disinformation detection in social media. Section 3 details the data collection and annotation processes. Section 4 presents the emotion analysis conducted on the dataset. Section 5 explores the linguistic patterns using the LIWC dictionary. Section 6 discusses the analysis of audio variables in TikTok posts. Section 7 describes the models used for disinformation detection and their performance. Finally, Section 8 concludes the study and outlines directions for future research.

2 RELATED WORK

The spread of misinformation and disinformation on social media has garnered significant research attention due to its profound impact on public opinion, democratic processes, and public health (Wang et al., 2022). The viral nature of social media platforms facilitates the swift dissemination of both true and false news (Vosoughi et al., 2018).

A seminal study by Vosoughi et al. (2018) analyzed the diffusion patterns of true and false news on X from 2006 to 2017, revealing that false news spreads significantly farther, faster, deeper, and more broadly than true news, particularly in political contexts. This enhanced virality is attributed to the novelty and emotional content of false stories, which evoke fear, disgust, and surprise, thereby increasing user engagement. Recent research has found that outrage is a key emotion that facilitates the spread of misinformation particularly on X (McLoughlin et al., 2024). Complementing these findings, Allcott and Gentzkow (2017) examined the role of fake news in the 2016 US presidential election, highlighting its substantial reach and influence, especially for stories favoring specific political candidates. The study underscored that users are more inclined to believe fake news that aligns with their preexisting beliefs, particularly within ideologically segregated social networks.

Addressing the challenges posed by misinformation, researchers have proposed various interventions. Pennycook et al. (2019) introduced the concept of the implied truth effect, demonstrating that attaching warnings to a subset of fake news headlines can inadvertently increase the perceived accuracy of untagged false headlines, highlighting the complexity of designing effective misinformation mitigation strategies.

The COVID-19 pandemic further highlighted the dangers of misinformation, giving rise to an “infodemic.” Islam et al. (2020) conducted a global social media analysis to assess the impact of false information related to COVID-19, finding that misinformation about treatments, transmission, and mortality rates fueled public panic and hindered effective public health responses.

In the realm of fake news detection, Shu et al. (2017) provided a comprehensive review of text mining techniques employed to identify false news on social media, proposing a conceptual framework that leverages both content-based and context-based features to enhance detection accuracy.

Multimodal approaches utilize both textual and visual data to enhance fake news identification. For instance, Rashid et al. (2024) introduced MuAFaNI, a framework that combines text and image representations using RoBERTa and ResNet-50 models to assess news authenticity. Similarly, Qu et al. (2024) proposed QMFND, a quantum multimodal fusion-based model that integrates image and textual features through a quantum convolutional neural network, thereby improving detection accuracy. Further advancing the field, Lee et al. (2024) explored the emotional effects of multimodal disinformation, demonstrating that video-based misinformation significantly increases anxiety and misperceptions compared to text-only content, highlighting the enhanced persuasive power of multimodal disinformation and its implications for public trust and decision-making.

Other notable contributions include Kumar and Taylor (2024), which developed a multimodal fake news detection model incorporating textual and visual features alongside feature engineering techniques to capture the behavior of fake news propagators. Shan et al. (2024) employed similarity inference and adversarial networks to fuse textual and visual features, thereby enhancing detection accuracy on social media platforms. Additionally, Pan et al. (2024) and Luvembe et al. (2024) introduced models that leverage graph convolutional networks and complementary attention mechanisms to align and fuse multimodal features, achieving significant accuracy improvements on benchmark datasets.

The emergence of deepfakes has further complicated the disinformation landscape. Vaccari and Chadwick (2020) investigated the impact of synthetic political videos, revealing that deepfakes contribute to uncertainty and reduce trust in news sources, thereby undermining the integrity of public discourse.

While multimodal approaches represent a significant advancement by leveraging both textual and visual data, the application of multimodal techniques specifically to TikTok videos remains underexplored. This study aims to address this gap by focusing on the multimodal characteristics of TikTok content, initially leveraging transcriptions of video audio, with the intention to incorporate video and audio features in future work.

3 DATASETS

3.1 Data Extraction

The datasets used in this study were extracted from TikTok and X using crawlers provided by *Apify*[2]. For TikTok, the *clockworks/tiktok-scraper*[3] was employed, while for X, the *fastcrawler/Tweet-Fast-Scraper*[4] was utilized. Additionally, the datasets will be made public for research purposes to facilitate further studies and enhance transparency in disinformation research.

The data extraction process was conducted using keywords related to the DANA phenomenon and associated disinformation narratives. The keywords, summarized in Table 1, were carefully selected to capture content that potentially spread disinformation or expressed emotional reactions during the DANA phenomenon.

Table 1: Keywords Used for Data Extraction and Their Translations.

Keyword (Spanish)	**Translation (English)**
DANA conspiración	*DANA conspiracy*
DANA fallecidos ocultación	*DANA deceased concealment*
DANA engaño	*DANA deception*
DANA manipulación	*DANA manipulation*
DANA mentiras	*DANA lies*
DANA ataque climático	*DANA climate attack*
DANA manipular la verdad	*DANA manipulate the truth*
DANA desinformación	*DANA disinformation*
DANA falsedades	*DANA falsehoods*
DANA ocultación	*DANA concealment*
DANA Rubén Gisbert	*DANA Rubén Gisbert*
DANA Alvise Pérez	*DANA Alvise Pérez*
DANA Iker Jiménez	*DANA Iker Jiménez*
DANA Vito Quiles	*DANA Vito Quiles*
DANA Cruz Roja falsa ayuda	*DANA Cruz Roja fake aid*
DANA Bonaire cementerio	*DANA Bonaire cemetery*
DANA ayuda rechazada	*DANA rejected aid*
DANA pronóstico incorrecto	*DANA incorrect forecast*
DANA radar sin funcionar	*DANA radar not working*
DANA provocada presas	*DANA caused by dams*

The extraction focused on filtering posts that matched these keywords, aiming to collect content that potentially spread disinformation or expressed emotional reactions during the DANA phenomenon.

For TikTok videos, the transcription of spoken content was extracted using the Whisper-medium model[5], which is designed for efficient and accurate speech-to-text transcription. Additionally, text

[2] https://console.apify.com/

[3] https://apify.com/clockworks/tiktok-scraper

[4] https://apify.com/fastcrawler/tweet-fast-scraper/api/javascript

[5] https://huggingface.co/openai/whisper-medium

appearing in the videos was extracted using PaddleOCR[6], a state-of-the-art Optical Character Recognition (OCR) tool. These methods ensured the comprehensive collection of both spoken and visual textual content for analysis.

3.2 Bias Assessment

To assess potential biases introduced by keyword-based data collection, we applied the Weirdness Index (WI) (Poletto et al., 2020), which measures how much certain words in a dataset deviate from their expected frequencies in a reference corpus.

As a reference, we collected social media content using broad, non-targeted keywords: "DANA Valencia". This corpus captures general discussions on the DANA phenomenon, including neutral and informative posts, rather than those explicitly linked to disinformation. By comparing our dataset against this reference, we evaluated whether our keyword selection introduced thematic biases.

Table 2 presents the WI statistics for both datasets. A high WI indicates that certain words are disproportionately represented compared to the reference corpus.

Table 2: Weirdness Index (WI) results comparing datasets.

Metric	Tweets	TikToks
WI Global (Mean)	4.86	1.76
WI Global (Median)	2.37	0.79
WI Standard Deviation	45.63	3.23
Words with WI > 2	55.64%	24.90%

The tweets dataset shows a significantly higher WI, indicating that keyword filtering has amplified specific narratives. Given the constraints in data availability and manual annotation time, our objective was to capture a broad range of disinformation-related themes rather than general discussions on the DANA phenomenon. The keyword selection was designed to encompass various disinformation topics, including conspiracy theories, exaggerated impacts, and misleading emergency response claims.

In contrast, the TikTok dataset exhibits a lower WI, suggesting a more diverse linguistic distribution. Although the same keyword selection strategy was used, TikTok's transcription-based content structure results in a dataset that, while still containing disinformation, is less narrowly focused compared to tweets.

3.3 Data Annotation

The collected TikToks and tweets were further annotated to create a labeled subset for analysis. Each post was examined to determine its relation to the DANA phenomenon and the spread of disinformation. Fact-checking was conducted using the reputable websites *Maldita*[7], *Newtral*[8], and *EFE Verifica*[9].

The annotation process involved assigning one of two labels to each post:

- **Label 0.** Assigned to posts that discuss the DANA phenomenon without spreading disinformation or falsehoods. This category also includes posts that actively report or debunk disinformation .
- **Label 1.** Assigned to posts that directly spread disinformation or falsehoods about the DANA phenomenon.

Table 3 shows the distribution of labels for TikToks and tweets. Efforts were made to ensure a balanced representation between the two classes (0 and 1) to facilitate unbiased analysis.

Table 3: Label distribution for TikToks and Tweets.

Label	TikToks	Tweets	Total
0 (Trustworthy)	131	177	308
1 (Disinformation)	137	205	342
Total	**268**	**382**	**650**

To better illustrate the difference between disinformation and trustworthy content, Table 4 presents one example of each category. These examples highlight the type of content analyzed and the criteria used for labeling.

The word clouds in Figure 2 highlight that the vocabulary across both social networks is strikingly similar. On X, terms related to *trustworthy* often include *bulo* (hoax), *tragedia* (tragedy), *falso* (false), and *desinformación* (disinformation), alongside mentions of public figures such as Iker Jiménez[10], Vito Quiles[11], Rubén Gisbert[12], and Alvise Pérez[13]. These terms often appear in posts criticizing or exposing falsehoods. In *disinformation*, words such as *embalse* (reservoir), *presa* (dam), *HAARP*[14], *Sánchez* (Span-

[6] https://github.com/PaddlePaddle/PaddleOCR

[7] https://maldita.es

[8] https://newtral.es

[9] https://verifica.efe.com/

[10] Iker Jiménez is a journalist.

[11] Vito Quiles is a pseudojournalist.

[12] Rubén Gisbert is a pseudojournalist.

[13] Alvise Pérez is the leader of *Se Acabó La Fiesta* (*The Party is Over*).

[14] Alleged climate attacks generated from HAARP (High-Frequency Active Auroral Research Program).

ish Prime Minister Pedro Sánchez), *UME*[15], *demolido* (demolished), *cementerio* (graveyard), and *AEMET*[16] are frequently observed.

On TikTok, the focus of terms associated with *trustworthy* shifts towards community and natural disaster responses, with words such as *inundaciones* (floods), *voluntario* (volunteer), *fenómeno* (phenomenon), *tormenta* (storm), *ayudar* (help), and *zona* (area) standing out. Disinformation in this context is more closely linked to terms like *Cruz Roja* (Red Cross), *ropa* (clothing), *verdad* (truth), *gobierno* (government), *radar* (radar), and *cadáveres* (corpses).

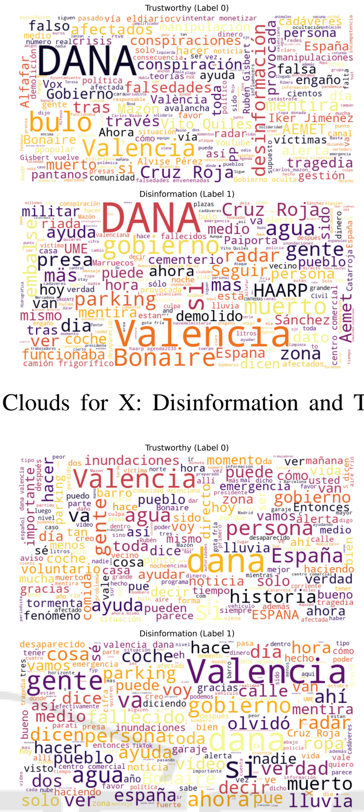
Table 4: Examples of Disinformation and Trustworthy Tweets (and their translation into English).

Trustworthy Tweet (Label 0)
Miles de voluntarios se movilizan desde Les Arts hacia las zonas afectadas. Matías Prats informa que no hay víctimas mortales en el túnel. (Thousands of volunteers are mobilizing from Les Arts towards the affected areas. Matías Prats reports that there are no mortal victims in the tunnel.)
Disinformation Tweet (Label 1)
Me acaban de comunicar que ya han sacado 86 fallecidos del parking de Bonaire. El aparcamiento del centro comercial tiene 5700 plazas. (I have just been informed that 86 deceased have already been removed from the Bonaire parking lot. The shopping center's parking lot has 5700 spaces.)

3.4 Few-Shot Annotation with GPT-4o

A Few-Shot annotation approach was tested using GPT-4o, a variant of GPT-4 optimized for enhanced efficiency and performance. While all annotations in this study were conducted manually, GPT-4o was applied to a subset of tweets to explore the potential of large language models for automated annotation.

Although the primary focus of our analysis was on two main categories (*Trustworthy information* and *Disinformation*), we discovered that refining the prompt to include four detailed categories and then collapsing them into two yielded higher agreement values between GPT-4o and manual annotations. This strategy allowed for a more nuanced understanding of the tweets during the initial annotation process, which improved the precision and reliability of the final binary labels.



(a) Word Clouds for X: Disinformation and Trustworthy Posts

(b) Word Clouds for TikTok: Disinformation and Trustworthy Posts

Figure 2: Word clouds from X and TikTok.

3.4.1 Prompt Design for GPT-4o

To classify the tweets into predefined categories, the following clean and structured prompt was provided to GPT-4o. For readers interested in the full details of the prompt structure, please refer to **Appendix A**.

3.4.2 Results and Evaluation

Table 5 shows the annotation agreement matrix for the four predefined categories, where GPT-4o achieved a Cohen's kappa of 0.684, indicating a substantial agreement with manual annotations. This highlights the effectiveness of GPT-4o in capturing nuanced differences between the categories.

Table 5: Annotation Agreement for Four Categories.

	0	**1**	**2**	**3**
0	15	1	0	2
1	1	13	1	1
2	3	4	28	5
3	4	2	3	36

When the categories were reduced to two groups,

[15] UME (Unidad Militar de Emergencias) is the Spanish Military Emergencies Unit.

[16] AEMET (Agencia Estatal de Meteorología) is the Spanish national meteorological agency.

Trustworthy (0) and *Disinformation (1)*, the Cohen's kappa further improved to 0.695 (Table 6).

Table 6: Annotation Agreement for Two Categories.

	0	**1**
0	66	8
1	9	36

These results demonstrate the power of GPT-4o, achieving substantial agreement with manual annotations. Such capabilities can significantly reduce the time and effort required for annotating large-scale datasets, particularly in the context of disinformation detection during events like DANA.

4 EMOTION ANALYSIS IN TWEETS AND TIKTOKS

This study analyzed the emotions present in tweets and TikToks using the *twitter-xlm-roberta-emotion-es* model[17]. The model identifies six emotions: *sadness*, *joy*, *anger*, *surprise*, *disgust*, *fear*, and *others*. These categories align with the basic emotions proposed by Paul Ekman, which are considered universal across cultures and rooted in human evolution Ekman (1992). Non-parametric tests (Mann–Whitney U) were performed to determine emotional patterns in disinformation across both X and TikTok.

4.1 Emotional Patterns in X

Table 7 presents the results for emotional analysis in tweets. Notably, tweets with disinformation showed significantly higher sadness (p = 0.000061). This often corresponds to claims exaggerating the number of casualties, e.g., "El engaño sobre el número de muertos es cruel, indignante e inhumano" ("The deception about the number of deaths is cruel, outrageous, and inhumane").

Disinformation also showed higher fear levels (p = 0.031), linked to claims about catastrophic events, such as dam failures, e.g., "Lo que pasa es que el embalse de Tous está al 97%. Si abren las compuertas, viene una riada" ("What's happening is that the Tous reservoir is at 97%. If they open the gates, a flood is coming").

Table 7: Emotional patterns in X.

Emotion	Mean (0)	Std (0)	Mean (1)	Std (1)	*p*-value
Sadness	0.077	0.131	0.176	0.235	0.000061
Fear	0.012	0.007	0.014	0.008	0.031030
Others	0.622	0.296	0.522	0.309	0.002906

4.2 Emotional Patterns in TikToks

Table 8 shows emotional patterns in TikToks. Unlike X, sadness is more prevalent in TikToks without disinformation (p = 0.000380), often focusing on the tragedy itself, e.g., "Lamentablemente las grandes inundaciones que se están provocando en Valencia España han dejado hasta la fecha más de 2500 desaparecidos" ("Unfortunately, the severe floods happening in Valencia, Spain, have left over 2,500 people missing so far").

Disinformation is associated with higher levels of anger (p = 0.000003) and disgust (p = 0.000003), reflecting frustration or exaggeration, e.g., "Las ayudas de los 6.000 euros para los afectados por la DANA las tienes que devolver en tres meses" ("The €6,000 aid for those affected by DANA must be repaid within three months"), or "Luego digan que no hay muertos en las calles, señor. Que no hay muertos en las calles. Acaban de tapar uno aquí con una toalla. Después de cuatro días recogiendo y amontonando cosas entre la mierda. ¿Eh? Y no había muertos. ¿Eh? No habían muertos" ("Then they say there are no dead bodies in the streets, sir. That there are no dead bodies in the streets. They just covered one here with a towel. After four days of picking up and piling things among the filth. Huh? And there were no dead bodies. Huh? There were no dead bodies").

Table 8: Emotional patterns in TikToks.

Emotion	Mean (0)	Std (0)	Mean (1)	Std (1)	*p*-value
Sadness	0.294	0.196	0.218	0.191	0.000380
Anger	0.165	0.143	0.262	0.187	0.000003
Disgust	0.027	0.018	0.040	0.025	0.000003

5 LINGUISTIC ANALYSIS USING LIWC

In addition to emotional analysis, linguistic patterns were examined using the LIWC dictionary Pennebaker et al. (1999)[18], which groups words into dimensions such as *Linguistic Dimensions*, *Psychological Processes*, *Personal Concerns*, and *Spoken Categories*.

[17] https://huggingface.co/daveni/twitter-xlm-roberta-emotion-es

[18] https://www.liwc.app/

5.1 Linguistic Patterns in TikToks

Table 9 summarizes the linguistic patterns in TikToks. When disinformation is not present, the discourse tends to be more articulate, with a higher use of prepositions and terms related to health, often informing about infection risks or providing warnings e.g., "El riesgo de infección en las zonas afectadas por la DANA sigue siendo muy alto debido a las aguas estancadas y contaminadas..." ("The risk of infection in areas affected by the DANA remains very high due to stagnant and contaminated water..."). Such videos also include more references to time and specific events e.g., "¡Atención! Una nueva DANA traerá humedad y precipitaciones a partir del martes por la noche" ("Warning! A new DANA will bring humidity and rainfall starting Tuesday night").

Conversely, videos spreading disinformation often use more negation and profanities (Maldec), reflecting frustration or skepticism. These videos also employ more perceptual words (Percept) to suggest first-hand observations or unverifiable claims e.g., "En las noticias no sale toda la verdad, están desaguando el centro comercial de aquí dicen que hay más de 800 me ha dicho un guardia civil..." ("The news isn't telling the whole truth; they're draining the mall here, and they say there are over 800 [bodies], a Civil Guard told me...").

Table 9: Linguistic Patterns in TikToks Using LIWC.

Category	Mean (0)	Std (0)	Mean (1)	Std (1)	*p*-value
Prepos	**10.68**	3.66	8.61	3.66	0.00001
Health	**0.25**	0.42	0.18	0.36	0.0424
Time	**2.56**	2.74	2.01	1.64	0.0261
Negation	1.37	1.28	**1.93**	2.58	0.0105
Maldec	0.11	0.30	**0.20**	0.42	0.0130
Percept	2.02	1.36	**2.56**	1.94	0.0336

5.2 Linguistic Patterns in X

Table 10 highlights the linguistic patterns observed in tweets. Similar to TikToks, tweets without disinformation also employ a more articulate discourse, using more prepositions (p = 0.000001) and negation (p = 0.036797).

In tweets spreading disinformation, there is a notable increase in the use of first-person singular pronouns (I). These tweets often cite unverifiable family connections as sources of information (Family), such as "Repito, tengo un familiar directo militar en el parking de Bonaire y están sacando muertos sin parar" ("I repeat, I have a direct family member in the military at the Bonaire parking lot, and they are removing dead bodies non-stop").

Table 10: Linguistic Patterns in X Using LIWC.

Category	Mean (0)	Std (0)	Mean (1)	Std (1)	*p*-value
Prepos	**14.08**	5.73	11.06	6.56	0.000001
Negation	0.86	1.67	**1.32**	2.21	0.0368
I	0.31	1.14	**0.54**	1.42	0.0403
Family	0.02	0.21	**0.13**	0.79	0.0320

The results indicate clear differences in linguistic patterns between disinformation and trustworthy content on both platforms. TikToks and tweets without disinformation tend to use more articulate language with a higher focus on factual information, while disinformation often incorporates perceptual language, negations, and personal anecdotes to enhance credibility

6 AUDIO FEATURES IN TIKTOKS

Audio is a key component of TikTok videos, contributing significantly to their emotional and persuasive impact. In this study, we extracted a wide range of audio features to analyze the role of sound in differentiating disinformation and trustworthy content. The extracted features include:

- **Zero Crossing Rate (ZCR).** It measures the rate at which the signal changes its sign, reflecting the noisiness or percussive nature of the sound.
- **Root Mean Square Energy (RMS).** It captures the loudness or energy of the audio signal, often linked to the intensity of the content.
- **Spectral Centroid.** It indicates the center of gravity of the spectrum, associated with the perceived brightness of the audio.
- **Spectral Rolloff.** It represents the frequency below which a certain percentage (e.g., 85%) of the spectral energy is concentrated, often used to distinguish harmonic and percussive elements.
- **Chroma Features.** They measure the energy distribution among the 12 pitch classes, representing harmonic content.
- **Spectral Bandwidth.** It quantifies the range of frequencies present in the signal, reflecting its tonal richness.
- **Spectral Flatness.** It describes the noisiness of the signal by measuring how flat the power spectrum is.
- **Mel-Frequency Cepstral Coefficients (MFCCs).** They capture the timbral aspects of the audio, essential for distinguishing speech and music characteristics.
- **Harmonics-to-Noise Ratio (HNR).** It indicates voice quality and clarity by comparing harmonic and noise components.

- **Tempo.** It estimates the speed of the audio, which can evoke urgency or calmness in the listener.
- **Onset Strength.** It measures the intensity of sound onsets, reflecting rhythmic and dynamic variations.
- **Pitch.** It refers to the perceived fundamental frequency of the signal, often linked to emotional tones such as excitement or seriousness.
- **Spectral Contrast.** It represents differences in amplitude between peaks and valleys in the power spectrum, providing insights into the dynamic range of the audio.
- **Tonnetz Features.** They capture tonal harmonic relationships, offering information about the musical nature of the audio.

For each feature, both the **mean** and **standard deviation (std)** were calculated to capture the central tendency and variability of the audio properties.

The analysis revealed significant differences in audio features between trustworthy (0) and disinformation (1) groups, as shown in Table 11, reflecting distinct acoustic and tonal patterns:

- **Zero Crossing Rate (ZCR, ZCR_STD).** Trustworthy audios exhibited higher values, indicating more frequent transitions across the zero axis in the waveform. This reflects more energetic and articulated speech, characteristic of clear and direct communication.
- **Spectral Features:**
  - *Spectral Centroid.* Higher in trustworthy content, representing brighter tones with more emphasis on higher frequencies.
  - *Spectral Bandwidth.* Wider in trustworthy content, indicating a broader range of frequencies around the centroid.
  - *Spectral Rolloff.* Higher in trustworthy posts, reflecting stronger energy in higher frequencies, which enhances tonal clarity.
- **Spectral Flatness.** Through the visualization of videos with high spectral flatness values, we observed that they frequently featured narrators delivering information in a neutral, monotone, or robotic style, often associated with objective news reporting about tragedies or serious events. This explains why spectral flatness was higher in trustworthy audios. In contrast, disinformation audios displayed lower spectral flatness, indicating more tonal content and greater variation. These variations are likely aimed at dramatizing events, evoking emotional responses, and manipulating the listener's perception.

Table 11: Significant Audio feature comparison between trustworthy (0) and disinformation (1) groups.

Feature	Mean (0)	Std (0)	Mean (1)	Std (1)	*p*-value
zcr	**0.094**	0.031	0.085	0.029	0.019
zcr std	**0.072**	0.030	0.060	0.029	0.006
spectral centroid	**1902.41**	465.65	1682.81	440.07	0.0002
spectral centroid std	**935.15**	313.20	798.22	337.14	0.0049
spectral rolloff	**3692.10**	996.91	3244.93	918.65	0.0010
spectral rolloff std	**1764.06**	470.15	1570.43	583.68	0.019
spectral bandwidth	**1984.54**	339.32	1799.34	348.43	0.0001
spectral flatness	**0.022**	0.029	0.017	0.028	0.007
spectral flatness std	**0.048**	0.063	0.044	0.069	0.030
onset strength	**1.80**	0.36	1.66	0.40	0.006
mfcc2 mean	111.10	25.04	**120.40**	28.08	0.022
mfcc5 mean	**3.16**	10.85	-0.44	11.02	0.010
mfcc6 std	**16.56**	3.77	15.54	3.97	0.021
contrast7 mean	49.03	2.60	**50.13**	3.67	0.025
contrast6 std	**7.58**	1.55	7.15	1.47	0.015
tonnetz4 mean	-0.001	0.051	**0.022**	0.074	0.003

- **Onset Strength.** Higher onset strength was observed in trustworthy audios, reflecting strong and clear acoustic events such as sharp changes in amplitude or the beginning of words and sounds. In contrast, lower onset strength in disinformation audios suggests smoother transitions and less dynamic emphasis, which might be used to create a more conversational or emotionally manipulative tone.
- **MFCCs (Mel-Frequency Cepstral Coefficients):**
  - *MFCC2.* Higher in disinformation, capturing lower frequency tonal characteristics, associated with deeper and emotionally charged tones.
  - *MFCC5 and MFCC6.* Higher in trustworthy content, indicating richer and clearer tonal

qualities in the mid-frequency range, aligned with neutral and structured speech.

- **Spectral Contrast (Contrast7).** Higher in disinformation, particularly in the highest frequency bands, helping dramatize content and attract attention.
- **Tonnetz (Tonnetz4).** Videos with high Tonnetz values often included eerie or mysterious music. Observational analysis of such videos revealed their use in disinformation to create a suspenseful or manipulative tone, amplifying the emotional impact.

Trustworthy audios are characterized by brighter tones, dynamic events, and monotone or robotic narration styles, optimized for clarity and credibility. Disinformation audios, on the other hand, emphasize tonal variation, emotional depth, and harmonic complexity, often accompanied by manipulative musical elements designed to captivate and influence the listener.

7 MODELS

To train the models for disinformation detection, tweets and transcriptions of TikToks were combined into a unified dataset. To minimize bias, hashtags and user mentions were removed from the text to prevent the models from relying on these features for predictions. Given the limited size of the dataset, we employed three main approaches: a traditional machine learning model, fine-tuning of transformer-based models, and a GPT-4o Few-Shot learning approach.

The models used were **roberta-base-bne**[19], **roberta-large-bne**[20], **xlm-roberta-large-twitter**[21], **longformer-base-4096-bne-es**[22], **fake-news-detection-spanish**[23], **SVM+TF-IDF**, and **GPT-4o Few-Shot**. We also experimented with fine-tuning **roberta-large-bne** by incorporating **audio features** from TikTok. These features were normalized and combined with text embeddings from the model's [CLS] token. The fused representations were passed through a classifier, leveraging both textual and audio information to enhance disinformation detection in a multimodal context.

To ensure robust evaluation, 5-fold stratified cross-validation was employed for the traditional and transformer-based models, maintaining the class distribution in each fold. This approach provided a comprehensive assessment of model performance, minimizing the risk of overfitting and ensuring reliable generalization to unseen data. For the GPT-4o Few-Shot approach, a single evaluation was conducted.

The performance of each model was evaluated using the following metrics: Accuracy, F1-Score, Precision, and Recall. The results are summarized in Table 12, where the best-performing model for each metric is highlighted in **bold**.

The *SVM+TF-IDF* model demonstrated the highest effectiveness in terms of *F1-Score* and *Recall*, achieving scores of 0.7949 and 0.8831, respectively. This superiority can be attributed to the Support Vector Machine's ability to efficiently handle high-dimensional feature spaces even with small datasets, thereby avoiding overfitting and providing robust generalization. The simplicity and efficiency of traditional models like SVM make them particularly suitable when data resources are limited.

On the other hand, transformer-based models such as *roberta-base-bne* and *roberta-large-bne* exhibited competitive performance, albeit slightly inferior to SVM in this scenario of limited data. However, it is important to note that these models possess significant potential for scalability and improved performance when larger training datasets are available.

Notably, by incorporating audio features into *roberta-large-bne*, we observed a significant improvement in both *Accuracy* and *F1-Score*, achieving 0.7646 in accuracy (an increase of 1.34 percentage points over the standard *roberta-large-bne*) and 0.7664 in F1-Score (an increase of 1.44 percentage points). Furthermore, this multimodal approach outperformed the *SVM+TF-IDF* model in terms of *Accuracy*, demonstrating the potential of combining textual and audio features to enhance performance in disinformation detection tasks.

Finally, the Few-Shot approach using **GPT-4o** achieved the highest accuracy of 0.7834, demonstrating the effectiveness of large-scale language models when provided with well-designed and specific prompts. This result highlights the potential of LLMs for automated annotation and classification tasks, even without domain-specific training.

[19] https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne

[20] https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne

[21] https://huggingface.co/sdadas/xlm-roberta-large-twitter

[22] https://huggingface.co/PlanTL-GOB-ES/longformer-base-4096-bne-es

[23] https://huggingface.co/Narrativaai/fake-news-detection-spanish

Table 12: Performance Metrics of Different Models for Disinformation Detection.

Model	Accuracy	F1-Score	Precision	Recall
roberta-base-bne	0.7497 ± 0.0506	0.7565 ± 0.0575	0.7668 ± 0.0313	0.7486 ± 0.0839
roberta-large-bne	0.7512 ± 0.0177	0.7520 ± 0.0270	0.7885 ± 0.0248	0.7220 ± 0.0543
roberta-large-bne + audio features	0.7646 ± 0.0292	0.7664 ± 0.0272	0.7726 ± 0.0432	0.7634 ± 0.0484
xlm-roberta-large-twitter	0.7274 ± 0.0175	0.7358 ± 0.0192	0.7505 ± 0.0167	0.7229 ± 0.0352
longformer-base-4096-bne-es	0.7266 ± 0.0473	0.7035 ± 0.0681	0.8069 ± 0.0361	0.6285 ± 0.0943
fake-news-detection-spanish	0.7298 ± 0.0454	0.7377 ± 0.0416	0.7541 ± 0.0486	0.7222 ± 0.0370
SVM+TF-IDF	0.7604 ± 0.0173	**0.7949** ± 0.0133	0.7227 ± 0.0165	**0.8831** ± 0.0087
GPT-4o Few-Shot	**0.7834**	0.7867	**0.8150**	0.7602

8 CONCLUSIONS AND FUTURE WORK

8.1 Conclusions

This study analyzed the spread of disinformation on TikTok and X during the DANA event in Valencia, Spain, which was marked by severe flooding. A dataset of 650 posts was manually annotated to distinguish between disinformation and trustworthy content. Additionally, a separate experiment was conducted using GPT-4o in a Few-Shot setup to evaluate its potential as a tool for automated annotation. GPT-4o demonstrated substantial agreement with the manual labels (Cohen's kappa of 0.684 for four categories and 0.695 for two categories).

Emotion analysis revealed that disinformation on X is primarily associated with heightened sadness and fear, while on TikTok, it correlates with increased anger and disgust. The identification of this divergence in predominant emotions across different social networks is a distinctive result of our study. First, our work reinforces the conclusion, already reached by many previous studies, that appealing to emotions is a recurring strategy in deliberately misleading messages Osmundsen et al. (2021); McLoughlin et al. (2024). However, thanks to the comparative approach of our study—where we analyzed the prevailing emotional traits in posts on X and TikTok—we have identified a phenomenon that has been scarcely investigated until now: the fact that, on each platform, disinformation content operates under a different emotional paradigm. This conclusion underscores the need to further understand the impact of emotional dimensions on credulity toward disinformation content, as well as the predisposition to share it Pennycook and Rand (2021). Linguistic analysis using the LIWC dictionary indicated that trustworthy content tends to use more articulate and factual language, whereas disinformation employs negations, perceptual words, and personal anecdotes to enhance credibility.

The analysis of TikTok audio features revealed distinct patterns: trustworthy audio showed brighter tones, stronger onset strength, and higher spectral flatness, reflecting clarity and neutrality, while disinformation audio used tonal variations and deeper MFCCs to evoke emotions and manipulate perceptions. These findings highlight the value of audio analysis in multimodal disinformation detection.

In terms of detection models, the traditional SVM+TF-IDF model outperformed transformer-based models and the GPT-4o Few-Shot approach in *F1-Score* and *Recall*, benefiting from the limited dataset size. However, transformer-based models showed competitive performance and potential for scalability with larger datasets. Notably, by incorporating audio features into *roberta-large-bne*, we observed an improvement in both *Accuracy* and *F1-Score* compared to its text-only counterpart, as well as a higher accuracy than the SVM+TF-IDF model. This demonstrates the value of multimodal approaches in disinformation detection tasks. The GPT-4o Few-Shot model demonstrated high accuracy, highlighting the effectiveness of LLMs in automated classification tasks without extensive domain-specific training.

8.2 Future Work

Future research can expand on this study by:

- **Increasing Dataset Size.** Incorporate more tweets and TikToks, as well as additional social media platforms.
- **Multimodal Analysis.** Include images and videos to capture a broader range of disinformation strategies.
- **Emotional Divergences.** Examine how distinct emotional triggers influence disinformation dissemination on different social media platforms.
- **Advanced Models.** Explore sophisticated transformer architectures and ensemble methods for enhanced performance.

- **Cross-Linguistic Studies.** Analyze disinformation patterns across different languages and regions.
- **Network Analysis.** Investigate user behaviors and network structures that facilitate disinformation spread.
- **Refining Few-Shot Techniques.** Improve prompt designs and methodologies for large language models to reduce reliance on manual annotations.
- **Intervention Strategies.** Develop effective counter-disinformation measures based on identified emotional and linguistic patterns.
- **Intelligent Agents.** Design and implement intelligent agents capable of real-time detection and mitigation of disinformation. These agents could combine machine learning techniques with rule-based systems to analyze linguistic and emotional cues, detect malicious content, and take automated countermeasures. For instance, intelligent agents could:
  - Flag potential disinformation for review or further analysis.
  - Provide users with context or verified information to counter false claims.
  - Interact with social media algorithms to limit the spread of harmful content while promoting verified, accurate information.

Addressing these areas will enhance the detection and mitigation of disinformation on social media, thereby strengthening the integrity of information ecosystems during critical events.

ACKNOWLEDGEMENTS

This work was done in the framework of the Iberian Digital Media Observatory (IBERIFIER Plus), co-funded by the EC under the Call DIGITAL-2023-DEPLOY-04 (Grant 101158511), and of the Malicious Actors Profiling and Detection in Online Social Networks Through Artificial Intelligence (MARTINI) research project, funded by MCIN/AEI/10.13039/501100011033 and by NextGenerationEU/PRTR (Grant PCI2022-135008-2).

REFERENCES

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

European Commission (2019). *Tackling online disinformation*.

Islam, M., Sarkar, T., Khan, S., Mostofa Kamal, A., Hasan, S., Kabir, A., Yeasmin, D., Islam, M., Amin Chowdhury, K., Anwar, K., Chughtai, A., and Seale, H. (2020). Covid-19-related infodemic and its impact on public health: A global social media analysis. *American Journal of Tropical Medicine and Hygiene*, 103(4):1621–1629.

Kumar, A. and Taylor, J. W. (2024). Feature importance in the age of explainable ai: Case study of detecting fake news & misinformation via a multi-modal framework. *European Journal of Operational Research*, 317(2):401–413.

Lee, J., Hameleers, M., and Shin, S. Y. (2024). The emotional effects of multimodal disinformation: How multimodality, issue relevance, and anxiety affect misperceptions about the flu vaccine. *New Media & Society*, 26(12):6838–6860.

Luvembe, A. M., Li, W., Li, S., Liu, F., and Wu, X. (2024). Caf-odnn: Complementary attention fusion with optimized deep neural network for multimodal fake news detection. *Information Processing & Management*, 61(3):103653.

McLoughlin, K. L., Brady, W. J., Goolsbee, A., Kaiser, B., Klonick, K., and Crockett, M. J. (2024). Misinformation exploits outrage to spread online. *Science*, 386(6725):991–996.

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., and Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *American Political Science Review*, 115(3):999–1015.

Pan, Z., Mao, Y., Xiong, L., Pang, T., and Ping, P. (2024). Mfae: Multimodal fusion and alignment for entity-level disinformation detection. *Pattern Recognition Letters*, 184:59–65.

Pennebaker, J., Francis, M., and Booth, R. (1999). Linguistic inquiry and word count (liwc).

Pennycook, G., Bear, A., and Collins, E. (2019). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*.

Pennycook, G. and Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*, 25(5):388–402.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477 – 523.

Qu, Z., Meng, Y., Muhammad, G., and Tiwari, P. (2024). Qmfnd: A quantum multimodal fusion-based fake news detection model for social media. *Information Fusion*, 104:102172.

Rashid, J., Kim, J., and Masood, A. (2024). Unraveling the tangle of disinformation: A multimodal approach for fake news identification on social media. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1849–1853, New York, NY, USA. Association for Computing Machinery.

Shan, F., Sun, H., and Wang, M. (2024). Multimodal social media fake news detection based on similarity inference and adversarial networks. *Computers, Materials & Continua*, 79(1).

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. In *Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 799–804.

Vaccari, C. and Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1):2056305120903408.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wang, S., Su, F., Ye, L., and Jing, Y. (2022). Disinformation: A bibliometric review. *International Journal of Environmental Research and Public Health*, 19(24):16849.

Wardle, C. and Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe, Strasbourg.

A DETAILED FEW-SHOT ANNOTATION PROMPT FOR GPT-4o

The following prompt was originally provided in Spanish to align with the language of the data but for the sake of comprehension for non-Spanish speakers, it has been translated into English in this article:

Context: The Depresión Aislada en Niveles Altos (DANA) that affected Valencia and other areas of Spain in October 2024 caused severe flooding and the spread of disinformation. Common types of disinformation included:

- **False claims about dam openings or breaks**, such as Forata.
- **Conspiracy theories about climate manipulation** (e.g., HAARP, Morocco, geoengineering).
- **Rumors of hidden deaths or transport in refrigerated trucks** (e.g., parking at Bonaire).
- **False accusations of rejected aid or mismanagement of donations**, such as delays or destruction of aid by the Red Cross.
- **Exaggerated numbers of victims, floods, or negligence**.
- **Alarmist content about basic services** (e.g., water supply, healthcare disruptions).
- **Unfounded accusations against political figures or governments**.
- **Claims of Valencia's meteorological radar being non-functional** (when it was operational).
- **Rumors of censorship or media manipulation** by public figures (e.g., YouTube or Iker Jiménez).

Labeling Guidelines:

- **0 (Not related to DANA):** Tweets unrelated to DANA, even indirectly.
  Example: "Just had the best paella in Valencia. Highly recommend!"
- **1 (Related to DANA, but not disinformation):** Tweets discussing real events or consequences of DANA without spreading disinformation.
  Example: "Thousands of volunteers are helping victims of DANA in Valencia."
- **2 (Disinformation mentioned, but criticized):** Tweets mentioning disinformation about DANA to criticize or report it, including posts debunking or ridiculing falsehoods.
  Example: "Don't believe the fake news about hidden deaths in Bonaire; it has been debunked by official sources."
- **3 (Disinformation):** Tweets actively spreading false information, rumors, or conspiracy theories about DANA, or unfounded accusations against political figures.
  Example: "They're hiding the truth! Over 800 bodies were found in the Bonaire parking lot, and no one is reporting it."

Tweet: [TWEET HERE]
Return only the number: