A Multidimensional Analysis of Text for Automated Detection of Computational Propaganda in Twitter

by

Marco Emanuel Casavantes Moreno

A dissertation submitted in partial fulfillment of the requirements for the degree of

Ph.D. IN COMPUTER SCIENCE

Doctoral Advisors:

Dr. Manuel Montes-Y-Gómez,

INAOE, Mexico

Dr. Luis Carlos González Gurrola,

Universidad Autónoma de Chihuahua, Mexico

Dr. Alberto Barrón Cedeño,

Alma Mater Studiorum–Università di Bologna, Italy

March, 2025

Santa María de Tonantzintla, Puebla, CP 72840, Mexico.

Instituto Nacional de Astrofísica, Óptica y Electrónica

Abstract

The way we consume news has been transformed, with technological advancements allowing people to easily express their views to vast audiences, including political opinions. These opinions can enhance a richer public dialogue; however, they also possess the potential to elevate extreme ideas that seek to manipulate or skew political narratives for personal benefit or agendas. Social media is frequently praised for its ability to boost political involvement, to the extent that its role in the spread of misinformation has even sparked worries about its impact on democracy. The importance of propaganda spread via social media can be linked to its influence in political matters, representing a domain where political factions compete for influence and control.

In the past few years, there has been a noticeable surge in the volume of research studies focused on the detection of propaganda across various domains, reflecting a growing recognition of the significance and impact of propaganda in contemporary society.

In this research study, we aim to contribute to the ongoing expansion of academic research surrounding the phenomenon of propaganda distributed through social networks, while also acknowledging the importance of various contextual factors that significantly influence the expression of propaganda in these environments. To facilitate this goal, we introduce a novel corpus specifically centered on propaganda posted and spread on Twitter, which has been collected from a diverse array of news media accounts.

By leveraging this unique dataset, we are putting forth a classification approach that incorporates a multitude of contextual attributes, thereby enabling a more effective detection of propaganda, particularly in comparison to a baseline strategy that focuses solely on the textual content of the messages without considering a broader context.

We have carried out an evaluation of the performance of our proposed approach across multiple data collections to assess its capabilities. From our evaluations, we report that our approach consistently outperforms the baseline classifier, demonstrating its superior effectiveness in detecting propaganda. Our analyses provide insights into what kind of contributions different contexts bring when detecting propaganda. Remarkably, we have even managed to secure the highest rankings in an international workshop dedicated to propaganda detection, where we competed against a multitude of other participating methodologies, further validating the significance of our contributions to this field of study.

Resumen

La forma en que consumimos noticias se ha transformado gracias a los avances tecnológicos que permiten a las personas expresar fácilmente sus opiniones a un público amplio, incluyendo sus opiniones políticas. Estas opiniones pueden enriquecer el diálogo público; sin embargo, también tienen el potencial de impulsar ideas extremas que buscan manipular o distorsionar las narrativas políticas para beneficio propio. Las redes sociales son frecuentemente elogiadas por su capacidad para impulsar la participación política, hasta el punto de que su papel en la difusión de desinformación ha suscitado incluso preocupación por su impacto en la democracia. La importancia de la propaganda difundida a través de las redes sociales puede vincularse a su influencia en asuntos políticos, representando un ámbito donde las facciones políticas compiten por influencia y el control.

En los últimos años, se ha observado un notable aumento en el volumen de estudios de investigación centrados en la detección de propaganda en diversos ámbitos, lo que refleja un creciente reconocimiento de la importancia y el impacto de la propaganda en la sociedad contemporánea.

En este estudio de investigación, buscamos contribuir a la continua expansión de la investigación académica en torno al fenómeno de la propaganda distribuida a través de las redes sociales, reconociendo al mismo tiempo la importancia de diversos factores contextuales que influyen significativamente en la expresión de la propaganda en estos entornos. Para facilitar este objetivo, presentamos un novedoso corpus centrado específicamente en la propaganda publicada y difundida en Twitter, recopilado a partir de diversas cuentas de medios de comunicación.

Al aprovechar este conjunto de datos único, proponemos un enfoque de clasificación que incorpora una multitud de atributos contextuales, lo que permite una detección más eficaz de la propaganda, especialmente en comparación con una estrategia de línea base que se centra únicamente en el contenido textual de los mensajes sin considerar un contexto más amplio.

Hemos llevado a cabo una evaluación del rendimiento de nuestro enfoque propuesto en múltiples conjuntos de datos para evaluar sus capacidades. A partir de nuestras evaluaciones, reportamos que nuestro enfoque supera continuamente al clasificador de línea base, lo que demuestra su eficacia superior en la detección de propaganda. Nuestros análisis proporcionan información sobre las contribuciones de los diferentes atributos contextuales a la detección de propaganda. Cabe destacar que incluso logramos las mejores clasificaciones en un taller internacional dedicado a la detección de propaganda, donde competimos con numerosas otras metodologías participantes, lo que valida aún más la importancia de nuestras contribuciones a este campo de estudio.

Agradecimientos

Al Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT), y al Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), por su apoyo otorgado durante la realización de este trabajo.

A mis asesores, Dr. Manuel Montes y Gómez, Dr. Luis Carlos González Gurrola, y Dr. Luis Alberto Barrón Cedeño, por compartir conmigo sus valiosos conocimientos y experiencias, así como brindarme su guía, lo cual me ayudó a crecer tanto como investigador como persona.

A mis sinodales, Dr. Luis Villaseñor Pineda, Dr. Jesús Ariel Carrasco Ochoa, Dr. Aurelio López López, Dra. Delia Irazú Hernández Farías, y Dr. Arkaitz Zubiaga, por sus valiosos comentarios, retroalimentación y observaciones, los cuales han sido fundamentales para mejorar este trabajo.

A mi familia y amigos, por su invaluable apoyo moral y emocional, sin importar la distancia. Gracias a mi padre Manuel, a mi madre Blanca, a mi hermana Cony, a mi abuelo Manuel y a mi abuela Cuqui.

A mis compañeros del INAOE, por acompañarme durante mi estancia en el instituto.

Dedicatoria

A mi familia, por estar conmigo y apoyarme en todo momento.

A mi madre, Blanca, que siempre me ha dado su apoyo incondicional, alentado a perseguir mis sueños, e impulsado a perseverar ante todos los desafíos que he encontrado en mi camino. Gracias por todas las conversaciones diarias que hemos compartido. Aunque me encuentre lejos de casa, siempre estás presente en mi vida.

A mi querido perro Buzz, porque aunque ya no esté con nosotros, su recuerdo es suficiente para alegrar mi día.

A mis amigos Fabián y Josué, por su apoyo, por reunirnos siempre que la oportunidad lo permite, porque a pesar de que hemos tomado caminos distintos, sé que cuento con su amistad, lealtad y hermandad.

A mi asesor, el Dr. Manuel Montes y Gómez, porque no puedo imaginar a otra persona que hubiera hecho más amena, interesante y divertida mi experiencia. Gracias por compartir conmigo su conocimiento, pero sobre todo gracias por darme la oportunidad de trabajar a su lado y por creer en mí.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The influence of social networks is widely regarded as incredible when it comes to the magnitude, extent, and speed of expansion they achieve constantly, evolving into a phenomenon that appears everywhere in our current daily lives [1]. Unfortunately, research findings suggest that these platforms have the potential to serve as channels for the dissemination of harmful, deceptive, or manipulative information [2]. Within the categorization of such information lies the concept of *propaganda*, which can be defined as *"an expression of opinion or action by individuals or groups, deliberately designed to influence opinions or actions of other individuals or groups with reference to predetermined ends"* [3].

Propaganda is frequently linked to the dissemination of news articles and political campaigns through conventional media outlets (like newspapers or websites) that prioritize news as their primary content. Nonetheless, certain investigations have suggested that the sources of information that individuals turn to and engage with have undergone a transformation, leading social media to branch out from its conventional role as a source of entertainment to also function as an online news provider [4]. In this context, it is observed that the content shared on such platforms tends to be noticeably shorter in length, noisier yet simpler to

digest, allowing for the rapid propagation of messages to a vast audience within a matter of seconds. Social networks are often praised for their potential to enhance political engagement, so much so that their influence in exacerbating political polarization and the widespread dissemination of misinformation has even raised concerns regarding their potential impact on democracy [5]. In this domain, political entities, including both democratic and antidemocratic factions, compete for power and control [6]. This situation raises several concerns regarding the ease with which populations can be influenced or manipulated, and more alarmingly, the potential for such influence to be carried out with malicious intent. Let us consider, for instance, the volume of information that was spread during the *2016 US Presidential campaign*, aimed to smear the reputation of specific candidates, or the safety and health measures that were not handed properly at the peak of the *COVID-19* global infodemic due to the quantity of disinformation disguised as reliable news [7].

Propaganda detection as a computational task has surprisingly not been explored as thoroughly in comparison to other categories of information disorders, such as Fake News or Hoaxes [8]. Consequently, there exist numerous aspects within this domain that have been neglected or treated in isolation when developing detection strategies, including but not limited to bias levels, geographical origin and emotions evoked. Each of these contextual variables represents a distinct dimension or perspective that is linked to techniques of propaganda. For example, *Political bias* can manifest through *"Name calling or labeling"* and *"Slogans"*. *Geographical background* can relate to *"Flag-waving"*. *Emotions* play a vital role in techniques such as *"Loaded Language"* and *"Appeal to fear/prejudice"*. A possible link between propaganda techniques and types of tweets is the necessity of identifying whether a message is directed specifically at another account (perhaps to incite *"Name calling"* or *"Doubt"*), simply retweeted (as a form of *"Repetition"*), or citing other sources (conceivably in an attempt to *"Appeal to*

authority") (more information about propaganda techniques can be consulted in Section 2.4). These aspects can be crucial when trying to influence the course of a discussion or argument.

In today's world, there is a pressing need for automated tools designed to assist in combating the challenges posed by propagandistic content. The main goal of this research is to explore the concept of propaganda within a social network, making comparisons to traditional propaganda while developing customized strategies that correspond to the various forms and degrees this content takes on a social platform. Through our investigation, this study seeks to evaluate messages disseminated on Twitter (currently referred to as *"X"*) by media outlets that have been classified as either reliable or dubious based on their promotion of propaganda. To achieve this goal, we curate a novel corpus specifically focused on Twitter-based propaganda, which has been meticulously gathered from a wide range of news media accounts. Using this distinctive dataset, we propose a classification methodology that integrates numerous contextual factors, thus enhancing the efficacy of propaganda detection, especially when contrasted with a baseline method that exclusively focuses on the texts of the messages without accounting for a broader context.

1.1 Problem Statement

Propaganda can be spread from many different sources, social networks being one of them. The volume of text-based exchanges in social media have made human intervention approaches unfeasible, and recent decisions and rulings by regulatory authorities explicitly mention automatic systems as tools to help mitigate the spread of mischievous content [9], proving their high social relevance.
Shared tasks are being held online to tackle this challenge and research is published to test new algorithms and approaches. The problem is that most of this

research is focused on propaganda extracted exclusively from news articles. Because of the lack of resources and limitations of previous work, there is research that acknowledges the room for improvement and necessity of further research on this subject [10]. To better solve the detection of computational propaganda issue, further exploration outside the news articles scope is needed. Since every day the influence of social networks grows as they become the main means of disseminating information, including malicious news and data, the goal of this work is to conduct a multidimensional analysis of computational propaganda. One of the characteristics that we have identified as a challenge is the existence of resources within the domain of news articles. In the course of our research, we have posed the question of whether these resources can be not only beneficial but also potentially adapted in some manner to facilitate the development of propaganda detection systems that could be implemented within the environment of a social network.

Furthermore, the second challenge that we have detected in our investigation is related to the prevalent manner in which propaganda tends to be analyzed. It is often executed with a primary focus on the textual content of the messages, thereby neglecting to take into account other factors that are intrinsically linked to the dissemination of propaganda. In a study about computational propaganda and political big data, Bolsover and Howard [11] suggest that a clear drawback of research based on big-data platforms is its dependence on readily available information (e.g., join date of the poster, friend counter, number of followers, number of total posts, etc.). Nevertheless, factors like geographic location, religious beliefs, political preferences, gender, level of education, and other variables that are typically linked to social behaviors are nearly impossible to obtain from Twitter data. This results in a significantly limited understanding of how these elements influence the spread of computational propaganda [11]. Consequently, we examine the question of whether the detection of propaganda could be substantially

enhanced by incorporating a broader consideration of the context surrounding the messages being analyzed.

1.2 Research Questions

- How can the resources from the domain of news articles be used to detect computational propaganda in Twitter?
- What are the differences (in terms of topics covered, emotions evoked) in computational propaganda from tweets based on its context?
- How can contextual information of messages be incorporated to improve the effectiveness of propaganda detection in them?

1.3 Main Objective

To assess a model for a multidimensional analysis of computational propaganda in tweets, taking advantage of resources on news articles, and considering different types of context, allowing to significantly improve the efficacy of current approaches.

1.3.1 Specific objectives

- To create a new propaganda corpus by collecting a minimum of 200,000 tweets from both non-propagandist and propagandist news sources on social media.
- To evaluate the performance of contextual classifiers on a dataset segmented by contextual features (at least two classifiers per feature), and measuring

performance using accuracy, precision, recall, and F1 score, with the goal of identifying differences in propaganda based on the contextual features and these metrics.

- To enhance propaganda detection in a statistically significant manner by training a classifier that incorporates contextual features such as bias, country of origin, and emotions evoked by texts, evaluating performance using accuracy, precision, recall, and F1 score.

1.4 Summary of Contributions

- A new corpus of propaganda from Twitter, *Propitter*, with over 385k tweets and extended with *Political Bias*, *Temporal information*, *Affective information* and *Geographic origin* as contextual features.

- Some insights about propaganda from social media. Through a comprehensive analysis, this study identifies differences in propaganda depending of the contextual features associated with it. In particular, it highlights the role of context in the detection of propagandist tweets, which was previously underexplored in the existing body of related literature.

- An approach that combines the tweet content and multiple contextual features for a better detection of propaganda in tweets. Using *Propitter*, our context-aware classifier exhibited a relative improvement of 7.08% (F1-score in the propaganda class) over a baseline without context.

1.4.1 Academic Production

1. Casavantes, M., Montes-y-Gómez, M., González, L. C., & Barróñ-Cedeno, A. (2023, November). *Propitter: A Twitter Corpus for Computational Pro-*

paganda Detection. In Mexican International Conference on Artificial Intelligence. Springer (pp. 16-27). Cham: Springer Nature Switzerland

This article introduces *Propitter*, the Twitter propaganda corpus that we created in the course of this study. The content of this article is included in Chapter 4.

2. Casavantes, M., Montes-y-Gómez, M., Hernández-Farías, D. I., González-Gurrola, L. C., & Barrón-Cedeño, A. (2023, January). *PropaLTL at DIPROMATS: Incorporating Contextual Features with BERT's Auxiliary Input for Propaganda Detection on Tweets.* In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023.*

   This article describes our participation in the DIPROMATS 2023 workshop (a propaganda detection task organized in IberLEF 2023). *DIPROMATS* datasets contain propaganda from Twitter accounts of diplomats, ambassadors, and governmental entities [12]. Part of the content of this article is included in Chapter 6.

3. Casavantes, M., Montes-y-Gómez, M., Hernández-Farías, D. I., González-Gurrola, L. C., & Barrón-Cedeño, A. (2024, January). *PropaLTL at DIPROMATS 2024: Cross-lingual Data Augmentation for Propaganda Detection on Tweets.* In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), Valladolid, Spain, September, 2024.*

   This article describes our participation in the DIPROMATS 2024 workshop (a propaganda detection task organized in IberLEF 2024).

4. Casavantes, M., Aragón, M. E., González, L. C., & Montes-y-Gómez, M.

(2023, October). *Leveraging posts' and authors' metadata to spot several forms of abusive comments in twitter. Journal of Intelligent Information Systems, 61(2), 519-539.*

This article is about experiments conducted to improve the detection of multiple types of Hate Speech using contextual attributes of users and their posts on Twitter.

5. Casavantes, M., Montes-y-Gómez, M., Hernández-Farías, D. I., González-Gurrola, L. C., & Barróñ-Cedeno, A. (2024). *PropitterX : A Twitter-based Propaganda Corpus Extended with Multiple Contextual Features. Submitted and currently under review in Language Resources & Evaluation.*

   This article describes the extension of *Propitter* with contextual attributes, creating *PropitterX*, data sub-collections, and corresponding experiments. The content of this article is included in Chapter 5.

6. Casavantes, M., Montes-y-Gómez, M., Hernández-Farías, D. I., González-Gurrola, L. C., & Barróñ-Cedeno, A. (2024). *A Contextual-Aware Approach to Detect Propaganda by News Outlets in Twitter. Work in progress, with the intention of submitting it to IEEE Transactions on Computational Social Systems.*

   This article details how we added contextual features to BERT-based classifiers, and the experiments performed on *PropitterX* and *DIPROMATS*. The content of this article is included in Chapter 6.

7. Casavantes, M., Hernández-Farías, D. I., & Montes-y-Gómez, M. (2025). *Entre la Información y la Manipulación: Detectando Propaganda en Tuits. Submitted (December 2024) and accepted (February 2025) in the Komputer Sapiens journal.*

   This article summarizes in Spanish our findings on propaganda detection using contextual features in the *Propitter* corpus.

1.5 Document Outline

The remainder of this thesis is organized as follows:

Chapter 2 contains an overview about the theoretical background concepts that serve as the foundation for the experiments and analyses that follow.

Chapter 3 presents related work on computational propaganda detection, where we discuss the main contributions and shortcomings of previous studies.

Chapter 4 describes the construction stages of our propaganda dataset from Twitter, denoted as "*Propitter*".

Chapter 5 introduces 4 sub-collections from *Propitter* and experiments based on contextual features (*political bias, geographical origin, affective information, temporal splits*).

Chapter 6 describes a classification approach for propaganda detection that leverages both content of tweets and contextual features.

Chapter 7 ends with our conclusions, scope, limitations and future work.

Chapter 2

Background

2.1 Machine Learning Algorithms

Natural Language Processing (NLP) is a branch of computer science and *Artificial Intelligence* (AI) that employs machine learning techniques to allow computers to comprehend and interact using human language[1]. *Machine Learning* (ML) focuses on enabling computers to change or adjust their actions (like making predictions), ensuring that these actions become increasingly accurate by measuring how closely the selected actions reflect the correct ones [13]. The multidisciplinary nature of machine learning becomes apparent as it is inspired by concepts from neuroscience and biology, statistics, mathematics, and physics, allowing computers to learn.

Machine Learning systems can be classified into broad groups according to the amount and type of supervision they get during training. Some of these categories are: supervised learning, unsupervised learning, semi supervised learning, and reinforcement learning [14]. When we feed data and the desired solutions or labels to an algorithm, we are talking about supervised learning, and a typical

[1] https://www.ibm.com/topics/natural-language-processing

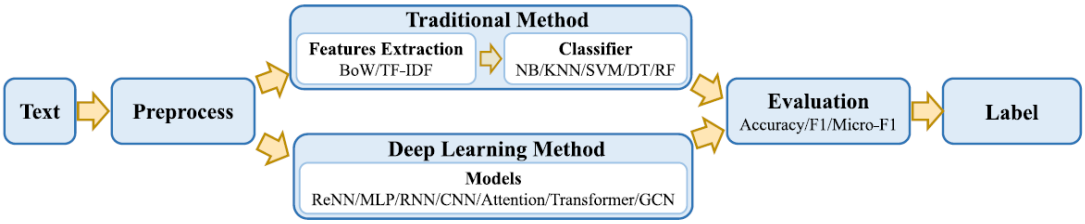

Figure 2.1: Text classification with conventional techniques in each segment. Identification of key features is vital for traditional approaches, whereas deep learning methods can automatically extract features. Flowchart adapted from [16].

task in this category is *classification*.

The classification problem consists of taking input vectors and deciding which of N classes they belong to, based on training from instances of each class. In one-class and multi-class classification problems, each example has one or more labels respectively, but for both tasks the set of classes covers the whole possible output space [13].

For this research, computational propaganda detection is treated as a text classification or categorization task, which is to assign a new document to one of a pre-existing set of document classes [15]. Text classification can be carried out under a traditional or a deep analysis (see Figure 2.1).

2.1.1 Traditional Method for Text Classification

Feature Extraction

Traditional Machine Learning uses a prominent feature representation to analyze and extract relevant insights from text data in NLP problems: *Bag-of-Words*. This representation model, commonly abbreviated as BoW, treats each word in a collection of documents as a feature, and since each document only contains a small subset of the whole vocabulary, BoW is an extremely sparse representation.

The value assigned to individual features can be either positive (if a given word exists within the document) or zero (if a given word is absent). The positive values can be term frequencies or simple binary indicators. For example, let us consider the next two documents:

- Doc1: "the weenie dog chases a cat"
- Doc2: "my cat likes dry food"

A BoW representation of these sentences, filled with binary indicators, would look like Table 2.1, where each column refers to a term and each row is a document.

Table 2.1: Example of a *Bag-of-Words*.

	the	**weenie**	**dog**	**chases**	**a**	**cat**	**my**	**likes**	**dry**	**food**
Doc1	1	1	1	1	1	1	0	0	0	0
Doc2	0	0	0	0	0	1	1	1	1	1

Alternatively, a BoW can also consider character n-grams (sequences of n number of items, in this case characters) as features (Table 2.2):

Table 2.2: Example of a Bag of Character 3-grams.

	the	**wee**	**een**	**eni**	**nie**	**dog**	**cha**	**has**	**ase**	**ses**	**cat**	**...**
Doc1	1	1	1	1	1	1	1	1	1	1	1	...
Doc2	0	0	0	0	0	0	0	0	0	0	1	...

There may be some applications (where a binary input is strictly required, or when presence is more important than frequency) for which binary representations are good enough. However, if frequency is indeed relevant for the task at hand, the use of frequencies of terms is a better way to fill the weights in the BoW. To achieve this, we attribute a weight to each term within a document, which is

determined by how frequently that term appears in the document. Our aim is to calculate a score that reflects the relationship between a term t and a document d, taking into account the weight of t in d. The most straightforward method is to set the weight equal to the number of times term t appears in document d. This method of assigning weights is known as *term frequency* [17].

Raw term frequency, as explained above, faces an issue: all terms are treated with equal importance. To mitigate the influence of terms that appear too frequently, a weighted variant is essential. For this reason, it is standard practice to use the document frequency df_t, which is defined as the number of documents within the collection that includes a term t.

How is the document frequency df of a term used to adjust its weight? Denoting the total number of documents in a collection as N, we define the inverse document frequency (idf) of a term t in the following manner:

$$idf_t = \log \frac{N}{df_t} \tag{2.1}$$

As stated in [17], we can combine the definitions of term frequency and inverse document frequency to create a combined weight for each term in each document.

The *tf-idf* weighting method assigns a weight to term t in document d represented by

$$tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t \tag{2.2}$$

In other words, $tf\text{-}idf_{t,d}$ provides a weight for term t in document d that is

1. at its peak when t appears frequently within a limited number of documents

(thereby giving those documents significant discriminating power);

2. diminished when the term appears less frequently in a document, or is found in numerous documents (thus providing a weaker relevance indication);
3. at its lowest when the term is present in nearly all documents.

To summarize the BoW model, the universe of words (or terms) corresponds to the dimensions (or features), turning them into a sparse multidimensional representation, where the ordering of the terms is not used.

Word Embeddings

Word ordering conveys semantics that cannot be inferred from the bag-of-words representation. For example, consider the following pair of sentences:

- "The cat chased the mouse"
- "The mouse chased the cat"

Clearly, the two sentences are very different but they are identical from the point of view of the bag-of-words representation. For longer segments of text, term frequency usually conveys sufficient evidence to robustly handle simple machine learning decisions. This is one of the reasons that sequential information is rarely used in simpler settings. On the other hand, more sophisticated applications with fine-grained nuances require a greater degree of linguistic intelligence. A common approach is to convert text sequences to multidimensional embeddings because of the wide availability of machine learning solutions for multidimensional data. However, the goal is to incorporate the sequential structure of the data within the embedding. Such embeddings can only be created with the use of sequencing information because of its semantic nature [18]. The simplest approach is to use a 2-gram embedding:

- For each pair of terms t_i and t_j the probability $P(t_j \mid t_i)$ that term t_j occurs just after t_i is computed.
- A matrix S is created in which S_{ij} is equal to $[P(t_i \mid t_j) + P(t_j \mid t_i)]/2$.
- Values of S_{ij} below a certain threshold are set to zero.
- The diagonal entries are set to be equal to the sum of the remaining entries in that row. This is done in order to ensure that the matrix is positive semi-definite.
- The top-k eigenvectors of this matrix can be used to generate a word embedding.

The linguistic power (semantic representation) in the embedding depends almost completely on the type of word-word similarity function that is leveraged [18].

The main idea behind this technique is that words that are similar in context (at least according to the text from which the embeddings algorithm trained with) appear closer to each other in a multidimensional space. Based on this, one can use the position of the words in this space to compute the similarity and relation that the text has with its surroundings.

Linear Support Vector Machine as Classifier

Introduced by Vapnik in 1992 [19], the Support Vector Machine (SVM) is a popular model in machine learning due to its versatility and power. This method works by finding the *support vectors*, the most useful data points in each class in a dataset that lie closest to a line called *classification line*. This line separates the classes in the best way possible (maximizing the margin or largest radius around it) before we hit a data point. This leads to an interesting feature of

these algorithms: after training this model we can throw away all data except for the support vectors, and use them for classification [13].

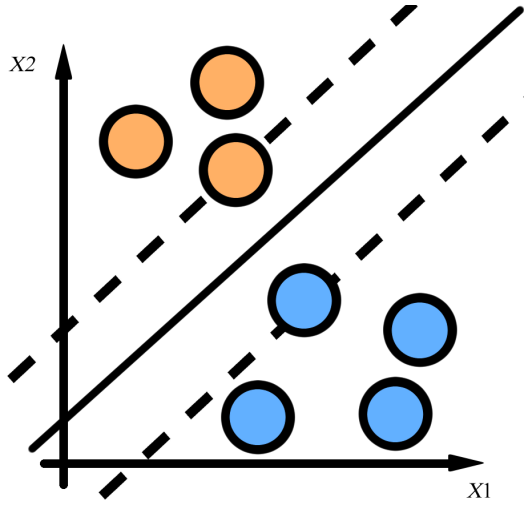

Figure 2.2: Example of a Linear Support Vector Machine. The solid diagonal line represents the classification line, while the dotted lines enclosing it represent the maximum margin between the classes.

In the "linear" version of the SVM (with a linear kernel), we can determine our classifier line by using the standard equation of the straight line:

$$y = w \cdot x + b \tag{2.3}$$

where w is the weight vector, x is the particular input vector, and b is the bias weight. For instance, we can use the classifier shown in Figure 2.2 by saying that any x value that gives a positive value for y is above the line and therefore an example of the orange class, and any x that gives a negative value becomes part of the blue class. A few distance constraints need to be added to take account of the margin. If we consider M to be the perpendicular distance between a dashed line and the classification line, we need to check if the absolute value of y is less

than M:

$$class(x) = \begin{cases} orange, & \text{if } y = w \cdot x + b \geq M \\ blue, & \text{if } y = w \cdot x + b \leq -M \end{cases}$$

However, this technique alone is not appropriate for datasets with outliers, since these kind of data points can make a classification problem non-linearly separable. In order to generalize and be useful for most real world cases, it needs to allow for some mistakes. This would be called a *Soft Margin Classifier*, as it has to look for the widest margin with the fewest classification mistakes, also named margin violations [14]. In a mathematical way, the function that we want to minimize is:

$$L(w, \epsilon) = w \times w + \lambda \sum_{i=1}^{R} \epsilon_i \tag{2.4}$$

where R is the number of misclassified points, and each ϵ_i is the distance to the correct boundary line for the missing point [13]. We can see that a new parameter is included, λ (which is also known as the C hyperparameter). A small λ means that we prioritize a large margin over a few errors, a large value of λ represents the opposite.

2.1.2 Deep Learning Method for Text Classification

Deep Neural Networks, commonly referred to as DNNs, are intricate systems that simulate the complex functionality of the human brain, enabling them to automatically learn and extract high-level features from data, and as a result, they often outperform traditional modeling techniques in various domains [16]. Depending on the specific characteristics of the data used, the corresponding input

word vectors are fed into the DNN for the purpose of training, and this process continues iteratively until a termination condition is satisfied. The effectiveness and performance of the training model are subsequently assessed and validated through various downstream tasks. These downstream tasks not only serve to evaluate the model's accuracy but also highlight the practical applicability of the DNN in real-world scenarios.

Pre-trained language models [20] are remarkable at grasping global semantic representations and elevate NLP tasks. They typically use unsupervised techniques to automatically discover semantic knowledge and then set up pre-training targets, allowing machines to learn how to comprehend semantics better [16].

As a pre-trained language model, the Transformer architecture relies on a comprehensive attention model and demonstrates efficacy in the domains of language, vision, and reinforcement learning, with its key component being the self-attention mechanism, which can be perceived as a graph-like induction bias that links all the tokens in a sequence through association-driven pooling operations [21].

Bidirectional Encoder Representations from Transformers

This representation technique better known as BERT by its initials, that can also be used to perform classification, solves a restriction that previous pre-trained language models had, unidirectional architectures. By masking a portion of tokens from the input in a random process called *"masked language model"*, a BERT representation is able to combine left and right contexts, generating a deep bidirectional Transformer. BERT's framework consists of a pre-training step, which involves training parameters on unlabeled data, and a fine-tuning step that continues adjusting these parameters, only this time with labeled data from downstream tasks. This process is illustrated in Figure 2.3 as a question-answering example.

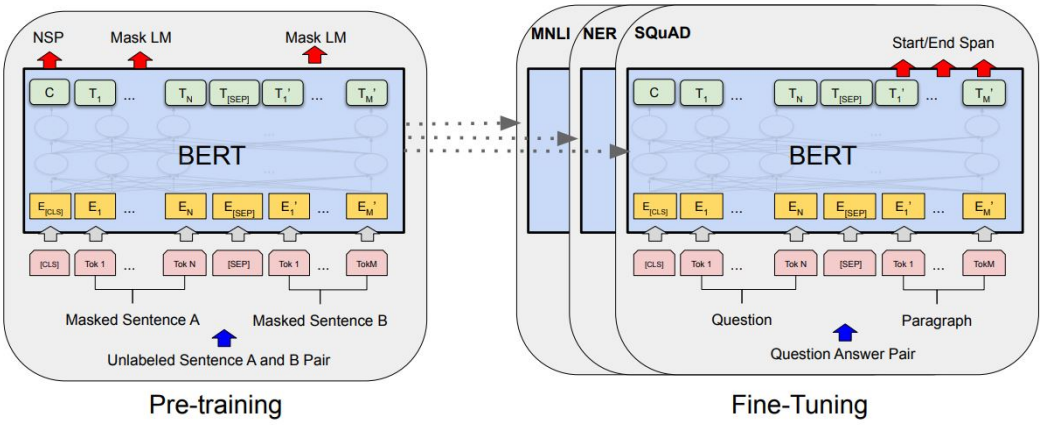

Figure 2.3: General pre-training and fine-tuning mechanisms in BERT, borrowed from [22]. Both pre-training and fine-tuning of parameters use the same architecture.

In BERT, a *"sentence"* refers to an arbitrary span of adjacent text, and a *"sequence"* indicates the input token sequence. Each sequence has special tokens, such as *"[CLS]"* which symbolizes the beginning of the input, and *"[SEP]"*, which separates sentences. The construction of an input representation for a given token, pictured in Figure 2.4, is the sum of the token, segment and position embeddings.

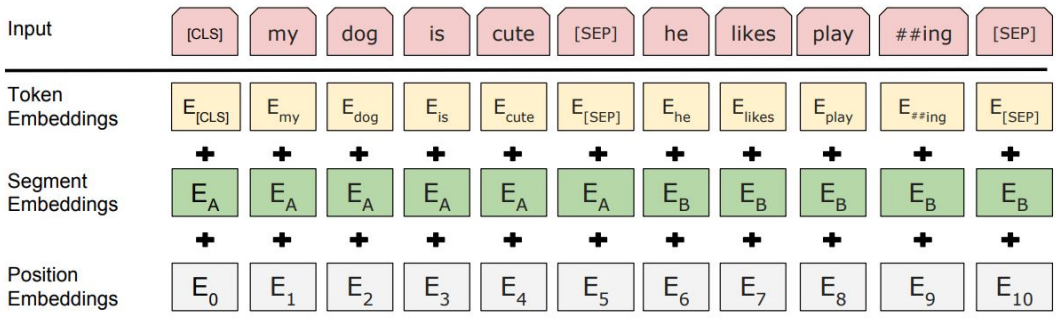

Figure 2.4: Example of BERT input representation, adopted from [22].

2.2 Evaluation measures

Classification tasks in supervised learning involves comparing predictions against the true labels of instances to train models. The possible outcomes of this comparison are shown in Figure 2.5.

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive (TP)**	**False positive (FP)**
	Predicted condition negative	**False negative (FN)**	**True negative (TN)**

Figure 2.5: The four outcomes of a confusion matrix.

The typical metric used to assess tasks such as classification is the F1 score, which serves as a balanced indicator of both precision (ratio of accurate positive results among all instances labeled as positive by a model) and recall (ratio of accurate positive results among all the actual positive instances in the data) in the detection process [23]. The F1 score is precisely the harmonic mean of these values. The calculations for precision, recall, and F1 are derived from the following terms:

- True positives (TP): the count of items that have been accurately assigned the class label;
- False positives (FP): the count of items that have been inaccurately assigned the class label; and
- False negatives (FN): the count of items that have been mistakenly labeled with a non-class label or a different class label.

Subsequently, the per-class precision, recall, and F1 score are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2.5}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.6}$$

$$F1\text{-}score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.7}$$

2.3 Types of information disorders

According to [24], there are two main kinds of information disorders based on the purpose behind it: on one hand we have *misinformation*, which includes unintentional falseness such as inaccurate dates, statistics or translations; and on the other hand there's *malinformation*, genuine information deliberately shared with an intent to harm, such as moving data intended for confidentiality into the open domain. A middle ground between these two exist in the form of *disinformation* (see Figure 2.6), intentionally false content created with the purpose of causing harm (false context, imposter, manipulated or fabricated content).

Not all propaganda content is generated with bad intentions. For example, there are cases in which propaganda is used to distribute positive messages such as raising awareness about the importance of voting, racial equity and the fight to promote women's rights [25]. It can also be used by content creators to find attractive ways to address news or events and catch the eye of potential readers. However, some other disorders fit inside the definition of it as a whole, such as

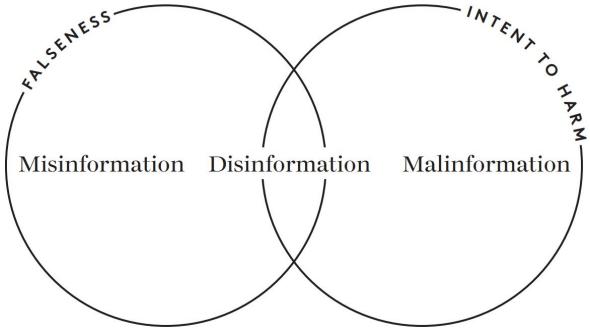

Figure 2.6: Types of information disorder, borrowed from [24].

Hoaxes or *Opinion Spamming* (see Figure 2.7) [2].

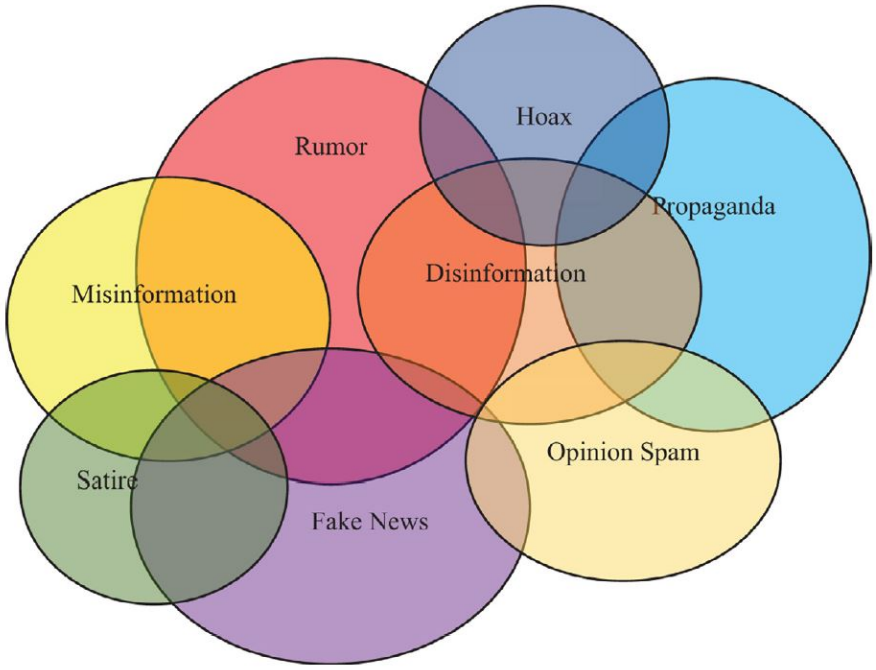

Figure 2.7: Venn diagram of false information on the Internet, borrowed from [2].

2.4 Propaganda techniques

The effectiveness of propaganda and misinformation is based on ideology and polarization [26], conveyed through different tactics. Clyde Miller, co-founder of

the Institute for Propaganda Analysis (IPA) [27] proposed in 1937 seven devices that appeal to emotions instead of reason [3]: *name-calling*, *glittering generalities*, *transfer*, *testimonial*, *plain folks*, *card stacking*, and *band wagon*. In 2019, Da San Martino et al. [28] listed the following 18 techniques:

1. **Loaded language.-** To affect an audience by using words and phrases with intense emotional connotations (either positive or negative).

   Example: *"The brave volunteers are risking everything to protect our children from the dangerous, radical forces trying to tear apart our community".*

   In this example, *"brave volunteers"* paints the individuals as heroic and selfless, *"risking everything"* adds urgency and sacrifice to their actions, *"dangerous, radical forces"* uses loaded terms to cast the opposition as extreme and harmful, and *"tear apart our community"* implies that the opposition threatens the very fabric of society, adding a sense of fear and urgency.

2. **Name calling or labeling.-** Using something the target audience either hates or loves to label the object of the propaganda campaign.

   Example: *"Those reckless communists want to destroy everything we've worked for. Don't let their radical agenda ruin our country."*

   In this example, *"reckless communists"* and *"radical agenda"* are used to label the opposition with negative, derogatory terms, painting them as dangerous and out of control.

3. **Repetition.-** Delivering the same message in a sustained manner until the audience finally accepts it.

   Example: *"A strong economy will make us a strong country and lead us to a strong future!"*

The repeated use of *"strong"* here reinforces the idea that strength is key to success.

4. **Exaggeration or Minimization.-** Representing something either in an exaggerated way or making it seem less important than it really is.

   Example: *"With this new policy, every family will have a perfect life—no poverty, no struggles, just endless prosperity!"*

   The exaggeration of the policy's potential impact creates unrealistic expectations and plays on people's hopes and desires.

5. **Doubt.-** To question the credibility of someone or something.

   Example: *"Can we really trust leaders who have failed us before? Are we sure they'll make the right choice this time?"*

   This creates doubt about the competence and reliability of the current leadership, making people question their past actions and motivations.

6. **Appeal to fear/prejudice.-** Seeking to support an idea by infusing anxiety and/or panic in a population towards an alternative, sometimes based on preconceived judgements.

   Example: *"If we don't pass this new security law now, our nation will be vulnerable to terrorist attacks. The enemy is already plotting against us, and without these measures, we could lose everything."*

   In this example, fear of terrorism and danger is used to persuade people to support the law, without addressing the actual merits of the law itself. The focus is on stoking fear to push for action.

7. **Flag-waving.-** Playing on intense national sentiment (regarding a particular group, such as race, gender, or political affiliation) to advocate for a specific action or idea.

Example: *"Our country, our pride! United we rise, divided we fall. Together, we are unstoppable!"*

This example aims to stir strong nationalistic feelings and a sense of unity among the people. The message appeals directly to the sense of ownership and pride citizens feel toward their nation, with a classic rallying cry that emphasizes the importance of unity.

8. **Causal oversimplification.-** Attributing a problem to a single cause when there are several factors contributing to it.

   Example: *"Crime is rising because our borders are weak. Close them, and our streets will be safe again."*

   This example reduces a complex issue *(rising crime)* to a single cause *(weak borders)* and implies that solving the simplified issue *(closing borders)* will immediately fix the problem. It ignores other potential contributing factors to crime, making the situation seem much more straightforward than it is.

9. **Slogans.-** A brief phrase that may include labeling and stereotyping.

   Example: *"Make America Great Again."*

   This slogan, widely used in political campaigns, invokes a sense of nostalgia and national pride, implying that a return to past greatness is possible if the right leader is chosen, while subtly casting doubt on current leadership or societal changes. The ultimate goal is to unite supporters under a vision of a restored ideal.

10. **Appeal to authority.-** Claiming that a statement is accurate solely based on the endorsement of a credible authority or expert in the field, without any additional evidence.

    Example: *"Experts agree this is the only solution for our future."*

This example uses the authority of *"experts"* (without specifying who they are or what their credentials are) to persuade people that a particular course of action is the right one. This technique relies on the implied trustworthiness of experts to make people accept the message without questioning it. It subtly suggests that dissent is unreasonable because the *"authoritative"* opinion has already been established.

11. **Black-and-white fallacy, dictatorship.-** Offering just two alternative choices as if they are the only options, despite the existence of additional possibilities.

    Example: *"You're either with us, or you're against us."*

    This slogan presents only two extreme options, ignoring any middle ground or nuance. It forces people to choose between two opposing sides, painting the situation as if there are no other alternatives, which simplifies decision-making but manipulates emotions and stifles critical thinking.

12. **Thought-terminating cliché.-** Expressions or terms that discourage critical thinking regarding a specific subject.

    Example: *"It is what it is."*

    By providing a simple, seemingly final explanation, this phrase discourages questioning or deeper analysis. Such clichés are used to stop people from challenging the narrative or considering alternative perspectives, effectively ending the conversation.

13. **Whataboutism.-** Undermine an adversary's stance by accusing them of hypocrisy without explicitly refuting their claims.

    Example: *"Why are you criticizing our government? <u>What about</u> all the corruption in other countries?"*

    This techinque deflects attention from the issue at hand by shifting focus

to a different, often unrelated problem. Instead of addressing the original criticism, it redirects the conversation, implying that because other nations might have worse issues, the current problem should be ignored or dismissed. It's used to avoid accountability and derail meaningful discussion.

14. **Reductio ad Hitlerum.-** Convincing an audience to reject a particular action or concept by indicating that it is favored by groups that the target audience despises.

    Example: *"You support this policy? Well, Hitler also believed in strong national borders, so you must be a Nazi."*

    This argument is dismissed by drawing a comparison to Hitler or Nazis, regardless of the actual merits of the policy or idea, in an attempt to demonize the position by associating it with a universally condemned figure, thereby avoiding any real discussion or analysis.

15. **Red herring.-** Introducing unrelated content to the topic at hand, causing the focus of everyone to shift away from the arguments presented.

    Example: *"We shouldn't worry about the government's new surveillance program. Think about how much we've improved our public transportation system! More buses and trains mean less traffic, fewer accidents, and cleaner air."*

    The message diverts attention from the controversial surveillance program (the real issue) by shifting the focus to unrelated improvements in public transportation, which does nothing to address concerns about privacy.

16. **Bandwagon.-** Seeking to convince the intended audience to participate and follow the same course of action because *"everyone around them is doing it"*.

Example: *"Everyone's switching to this new energy drink. It's the most popular choice among athletes and fitness enthusiasts! Don't get left behind—join the trend and try it for yourself!"*

This example encourages people to buy the energy drink simply because it's popular, implying that if everyone else is doing it, they should too, without offering any real reasons why it's a better product.

17. **Obfuscation, intentional vagueness, confusion.-** Employing intentionally vague language, so that the audience may have its own interpretation.

    Example: *"Our new program's holistic approach to optimizing fiscal efficiency aligns perfectly with our long-term strategic objectives, enhancing overall national prosperity."*

    In this example, the language is vague with words like *"holistic approach"*, *"optimizing fiscal efficiency"*, and *"long-term strategic objectives"*, which make the program sound very positive without actually explaining what it entails or how it will impact the average person. The lack of clear, specific information is meant to obscure the true nature of the program, leading the audience to trust it without questioning the details.

18. **Straw man.-** When an opponent's suggestion is replaced with a comparable one that is subsequently countered instead of the original.

    Example: *"Opponents of our healthcare reform argue that we shouldn't invest in better healthcare at all, claiming that it's a waste of taxpayer money. But we know that a healthier nation is a stronger nation, and rejecting improvements to our healthcare system would be irresponsible and harmful."*

    In this example, the opposition's argument is misrepresented as being against all healthcare investment, which is likely not their actual position. The strawman simplifies and distorts the argument to make it easier to

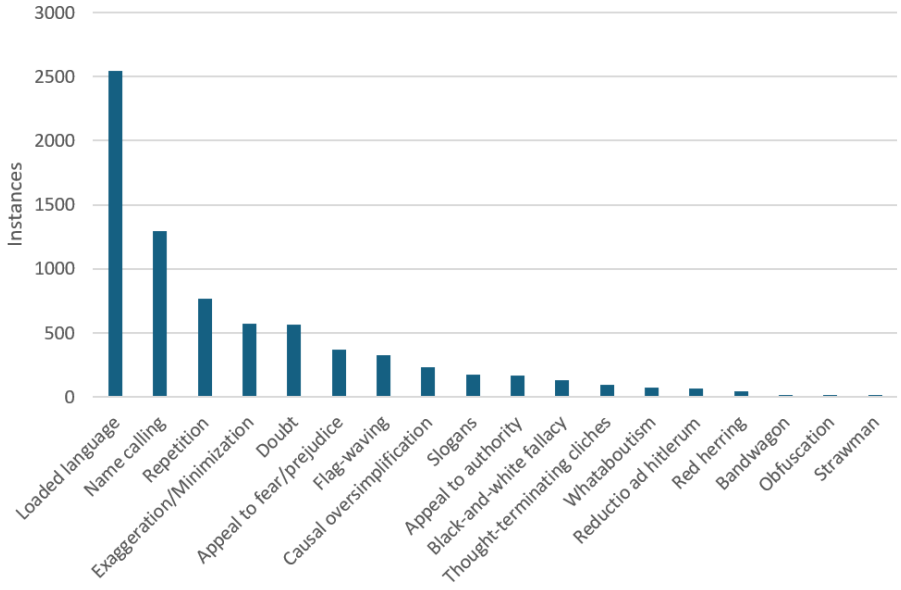

Figure 2.8: Frequency of propaganda techniques in the PTC corpus [28].

refute by framing it as a reckless stance, rather than addressing the real concerns or nuanced points raised by the opposition.

The Figure 2.8 shows an approximation of the most used propaganda techniques based on the frequency with which they appear in the PTC corpus from [28], with the five most popular techniques being:

1. Loaded language
2. Name calling
3. Repetition
4. Exaggeration or Minimization
5. Doubt

Chapter 3

Related work

The emergence of the Internet and social media has significantly altered the landscape of propaganda, enabling a broader range of individuals and groups to create and spread propaganda messages, a task that was previously exclusive to governments and major organizations [11]. Additionally, it has introduced new opportunities for the swift dissemination of propaganda, achieved through the manipulation of online information algorithms and processes, as well as the targeting of specific audiences using advanced data analysis techniques. Thus, in this section we divide the research focused on computational propaganda *outside* and *inside* social networks.

3.1 Computational Propaganda Detection Outside Social Networks

This particular section is dedicated to presenting an overview of the foundational research studies conducted in the area of computational propaganda detection. At the beginning of these efforts, the primary focus was on identifying propagandistic

Table 3.1: News articles in TSHP-17 corpus, adapted from [29].

News Type	Source	# of Documents
Trusted	Gigaword News	13,995
Propaganda	The Natural News	15,580
	Activist Report	17,869

content within various news articles, which required linguistic analyses of these documents in their entirety. Later studies followed in with a deconstruction of these documents into individual sentences for further investigation. The most noteworthy research contributions that have been made in this particular domain are summarized in the subsequent sections below.

3.1.1 Propaganda as part of Fake News Analysis

Propaganda detection can be conceptualized as a text classification task. A group of scholars conducted an investigation into the language employed by news media within the realm of political fact-checking and identification of fake news [29]. Their study involved a comparison of the linguistic features of authentic news content against those of satire, fraudulent information in the form of hoaxes, and propaganda, with the aim of identifying distinctive characteristics indicative of untrustworthy texts. In order to examine linguistic patterns among various genres of articles, a selection of trusted news articles from the "English Gigaword" corpus [30] (a large collection of newswire text data in English amassed by the Linguistic Data Consortium over the course of several years) was analyzed alongside articles retrieved from seven unreliable news websites spanning different categories. One of the categories explored was propaganda, defined as content designed to deceive readers into believing a specific political or social ideology. Table 3.1 shows the quantity of articles in the *TSHP-17* dataset introduced by [29].

The crawled articles were used for the purpose of *News Reliability Predic-*

tion, where a Max-Entropy classifier with L2 regularization was trained. This involved feeding the classifier with feature vectors consisting of unigrams, bigrams, and trigrams with a *tf-idf* weighting scheme.

As part of their research findings, they delineate that the most significant weighted n-grams associated with reliable news frequently refer to particular locations (e.g., *"washington"*) or temporal references (*"on monday"*), while prominently weighted characteristics indicative of propaganda lean towards conceptual generalizations (*"truth"*, *"freedom"*) along with specific topics (*"vaccines"*, *"syria"*) [29].

3.1.2 Propaganda Detection as a Standalone Task

Beginning with an examination of the shortcomings associated with the *TSHP-17* corpus, it lacks in providing information about the origins of each news article in it. Furthermore, it managed to gather information from a relatively small selection of sources, eleven in total, with only two of those sources being categorized as propagandistic. Consequently, conducting thorough experiments and analyses considering the source factor was unfeasible. These limitations prompted the development of a new corpus named *QProp* [31]. In that study, a binary class classification was conducted, starting to shape propaganda detection as a standalone task and distancing it further from the fake news scope. They considered 94 sources of *non-propaganda* and 10 sources of *propaganda* (Table 3.2 displays the distribution of their collection). The labels were produced by using the news sources as labeling mechanisms—commonly referred to as *distant supervision* from MediaBias/FactCheck[1], a website that categorizes media, journalists, and politicians. By increasing the size of their corpus selecting more propagandist news sources, systems trained with this data could learn to distinguish propaganda instead of learning the writing and publishing style of the

[1]https://mediabiasfactcheck.com/

Table 3.2: News articles in QProp, adapted from [31].

News Type	Sources	# of Documents
Trustworthy	94	45,557
Propagandistic	10	5,737

news outlets. Their hypothesis was that representations based on writing style and readability can generalize better than approaches based on word-level representations. For their experiments, they used a Max-Entropy classifier with L2 regularization, feeding it features based on word n-grams, lexicons, vocabulary, and *NELA* [32]. Their findings demonstrate that models capturing writing style and text complexity exhibit superior effectiveness compared to word n-grams.

3.1.3 Fine-Grained Analysis of Propaganda

In 2019, the *PTC corpus* was introduced [28], encompassing new features compared to those present in previous collections. This dataset was characterized by its manual annotation process, which stood in contrast to the conventional practice of employing news sources as *distant supervision*. Furthermore, this dataset was annotated at the *span level*, an advancement that involved identifying and marking specific snippets of text, thereby allowing for a more granular analysis rather than categorizing entire documents as a whole. The second twist of this research was the transition from a binary classification framework to a more complex multi-class classification scheme, which took into account a total of 18 distinct propaganda techniques, thereby enhancing the depth of the study in terms of propaganda analysis. Although there are some techniques that appear only a few times in the collection (e.g. *"straw man"*, with only 15 instances out of a total of 7,485), it is worth mentioning that *Loaded language* and *Name calling or labeling*, the two most popular techniques (appearing 3,841 times combined, more than half the instances in the whole collection) share an association with

Table 3.3: News articles in PTC, adapted from [28].

News Type	Sources	# of Docs.	Prop. Techniques	Instances
Non-propagandistic	36	79	–	–
Propagandistic	13	372	18	7,485

the use of emotions as a way to "push" propagandistic content into the messages.

As an interesting fact, the authors of *PTC corpus* now labeled the *"trustworthy"* class as *"non-propagandistic"*, perhaps as a result of the difference in task purpose between fake news and propaganda detection. Table 3.3 shows the distribution of the *PTC* corpus.

NLP4IF 2019

In 2019, the second workshop on NLP for Internet Freedom (NLP4IF)[2] presented two subtasks involving propaganda detection using the *PTC* corpus, one for identification of propagandist texts at the fragment-level and a binary classification task at the sentence-level [33]. In the Sentence-Level Classification in the Test Set, 9 out of 10 teams reported the use of BERT [22] in some form to predict labels, either independently or as part of an ensemble. Other teams from the top scores (shown in Table 3.4) found useful to consider lexical features, sentiments, and tackling the class imbalance of the set to achieve their final results. We can observe that all the top strategies proposed for this task were dependent on Transformer architectures. For the Sentence-Level Classification in the Development Set, the best performance was achieved by a combination of three classifiers [34]: two based on BERT [22] and one on Google's Universal Sentence Encoder [35].

[2] `http://www.netcopia.net/nlp4if/2019/index.html`

Table 3.4: Top Official Results for NLP4IF SLC Task - Test Set. Adapted from [33].

Rank	Classifier	F1	System Description
1	BERT	0.6323	Attention Transformer trained on Wikipedia and BookCorpus.
2	BERT	0.6249	Over-sampled training data and performed cost-sensitive classification.
3	BERT	0.6249	Ensemble of models.
4	BERT + LR + CNN	0.6230	Voting ensemble with features from FastText embeddings, readability, emotions and sentiments.
5	N/A	0.6183	Not reported.
6	BERT + USE	0.6138	Ensemble of two BERTs and Universal Sentence Encoder.
7	BERT + bi-LSTM + XGBoost	0.6112	Ensemble with features from GloVe embeddings, affective and lexical representations.

SemEval-2020 Task 11

Task 11 of SemEval-2020 focused on the detection of propaganda techniques in news articles [36], concentrating on fine-grained analysis of texts that could complement existing strategies. Again, practically all approaches submitted for this task relied on systems based on Transformers. The best ranked team for Span Identification [37] trained several of these architectures and combined them in the end as an ensemble. This result, along with the rest of participants among the top five teams, is displayed in Table 3.5.

3.1.4 Propaganda from Digital Newspapers and Web Pages

Polonijo et al. [38] introduced a deep learning technique to merge sentiment scores with Word2Vec [39] vectors, resulting in a representation that encompasses both semantic and emotional data, which leads to a more accurate model for propa-

Table 3.5: Top Results for SemEval-2020 Task 11 Span Identification - Test Set. Adapted from [36].

Rank	Classifier	F1	System Description
1	Ensemble of 6+ architectures	51.74	Complex heterogeneous multi-layer neural network with BIO encoding, Part-of-Speech and Named Entity embeddings.
2	RoBERTa	49.88	Ensemble of models with oversampling by producing silver data.
3	RoBERTa	49.59	Ensemble with attached CRF for sequence labeling.
4	BERT+ BiLSTM	48.16	Model with extra features (PoS, NE, sentiment) and fine-tuned on 10k additional propaganda articles.
5	BERT	46.63	Used masked language modeling to domain-adapt their base model with 9M articles (fake, suspicious, hyperpartisan news).

ganda classification. Word2Vec vectors serve as an effective tool for understanding the semantic significance of words, and an emotional lexicon incorporated into VADER's [40] sentiment analysis yields a sentiment score for the text that encapsulates emotional insights. This approach maintains the adaptability of the Word2Vec vector by fusing it with the outcomes of sentiment analysis. The data they analyzed consists of two parts: propaganda data from texts in English from the Xinhuanet[3] and CGTN[4] newspapers' Internet portals, and non-propaganda texts from news articles from Reuters[5] and TheHill[6]. The experiments they conducted using a Word2Vec model with sentiment data alongside standard deep learning techniques for propaganda detection from extracted text from web pages (comprising 37,503 lines of *propaganda* and 43,613 lines of *nonpropaganda* text) showed that their strategy enhances the accuracy of propaganda classification.

[3]http://www.xinhuanet.com/english/
[4]https://www.cgtn.com/
[5]https://www.reuters.com/
[6]https://thehill.com/

3.2 Computational Propaganda Detection in Social Networks

The research studies summarized below considered and evaluated forms of propaganda that have been disseminated across various social networking platforms, thereby highlighting the growing impact and influence that these digital communication channels apply on public opinion and societal discourse.

3.2.1 Propaganda Disseminated in Twitter

Wang et al. [10] explored propaganda from different sources. They hypothesize that propagandistic sources are sophisticated and creative, and that they will find new ways to deceive by evading trained classifiers. The novelty of their approach lies in *cross-domain learning*, recognizing the scarcity of labeled data where domains represent different types of sources, such as news articles, social media posts, and public speeches. The data collections used for their experiments fall into precisely these three types of sources. Table 3.6 shows distribution of these corpora. They created a collection of speech transcripts from four politicians, arranged in ordered pairs. *Trump* and *Obama* as contemporary speakers. *Trump* was seen as more propagandist than *Obama*. They also use *Joseph Gobbels* (Nazi Propaganda Minister) and *Winston Churchill* (UK Prime Minister) as important figures around the time of World War II, *Gobbels* supposedly being more propagandistic than *Churchill*. All four of these politicians have given propaganda speeches, and the author's supposition is that two of the speakers exhibit less propaganda than the other two.

With news as a source, they combined and reorganized the datasets used in *"Hack the News"*[7], to build an article-level corpus and a sentence-level corpus. With

[7] https://www.datasciencesociety.net/hack-news-datathon/

Table 3.6: Distribution of cross-domain corpora from [10]. *"+Prop"* and *"-Prop"* means a politician was considered more propagandistic or less propagandistic, respectively.

Source	Class	Documents	Sentences
Speeches	Trump (+Prop)	100	7,985
	Obama (-Prop)	100	8,336
	Goebbels (+Prop)	44	4,482
	Churchill (-Prop)	44	4,131
	TOTAL	288	24,934
News	Propagandistic	3,899	3,938
	Normal	3,899	3,938
	TOTAL	7,798	7,876
Tweets	Propagandistic	–	8,963
	Normal	–	8,963
	TOTAL	–	17,926

tweets as a source, they combined two collections, Twitter Internet Research Agency Dataset (*Twitter IRA*)[8] from 2018, and *twitter7* [41], a 2009 collection of almost 476 million tweets. They used the 8,963 tweets from the *Twitter IRA* as examples of propaganda, and an equal number of tweets extracted from *twitter7* as examples of normal instances.

The four propaganda detection methods that they used were divided in two types:

- Attribute-based models: Logistic Regression and Support Vector Machines. The features considered were word count, weighted n-grams with TF-IDF, and LIWC [42] word categories.
- Models based on neural networks, a *Long Short-Term Memory* (LSTM) baseline and a modification to this baseline, which is a contribution of this work that they call the LSTM or *Long Short-Term Memory Regressor*

[8] Available at `https://archive.org/details/twitter-ira`

(LSTMR) Pairwise Classification Model (they designed a model that relaxes the constraints of strict labeling on rankings).

As part of their analysis, they concluded that the best cross-domain results are obtained when training with news and applying those models to speeches or tweets, and that the cross-domain classification excluding names leads to poorer performance. Their findings also suggest that exaggerations (e.g. *"absolutely"*) and negative emotions (e.g. *"lies"* or *"devastating"*) play a key role in audience manipulation. Regarding the characteristics of LIWC, words that express negative emotions are typical of propaganda.

TWEETSPIN is a collection of tweets that feature weak labels indicative of propaganda techniques [43]. It contains 210,392 tweets with 19 labels, referring to 18 *propaganda* techniques and 1 *non-propaganda* label (see Table 3.7). The propagandist instances were selected by retrieving tweets containing keywords related to the propaganda techniques. Following this, they introduced MVPROP, a transformer-based model for multi-view propaganda detection, which assimilates multi-view contextual embeddings through pairwise cross-view transformers. They illustrated how enriching the input tweet text with semantic, relational, and knowledge views significantly enhances performance compared to other baseline approaches. Their experiments also confirmed the adaptability of their trained model in detecting propaganda within news articles.

As part of IberLEF 2023 [44], DIPROMATS was organized with the goal of identifying techniques to detect propagandistic tweets from governmental and diplomatic entities [45]. It introduced three subtasks across two languages, Spanish and English: i) A binary classification task to determine whether or not a tweet employs propaganda techniques, ii) A multiclass, multilabel classification task, where systems must ascertain, for each tweet, which of the 5 available categories (*Not propagandistic*, *Appeal to Commonality*, *Discrediting the Opponent*,

Table 3.7: Dataset statistics for TWEETSPIN, adapted from [43].

#Total Propaganda Tweets	157,327
#Total Non-Propaganda Tweets	53,165
Propaganda technique	**# Tweets**
Loaded Language	18,365
Name Calling/Labeling	17,096
Reductio Ad Hitlerium	15,677
Doubt	14,993
Appeal To Fear/Prejudice	14,654
Whataboutism	13,887
Repetition	13,285
Slogans	10,190
Appeal To Authority	8,539
Flag-Waving	7,675
Exaggeration, Minimization	5,416
Black-And-White Fallacy	4,872
Thought-terminating cliches	3,781
Bandwagon	2,547
Red Herring	2,315
Causal oversimplification	1,790
Straw man	1,265
Obfuscation, Intentional Vagueness, Confusion	1,048

Loaded Language. *Appeal to Authority*) it belongs to, and iii) A fine-grained classification task where systems need to identify which specific techniques are present in the tweet. The distribution of the DIPROMATS dataset is shown in Table 3.8. As this corpus was used in this study to conduct experiments, more information about its data distribution is provided in Section 6.3.4 (Table 6.6).

3.2.2 Propaganda Disseminated in Reddit

Balalau and Horincar examined how propaganda influences six prominent political forums on Reddit (see Table 3.9) that target a varied audience across two nations, the US and the UK [46]. They determined that the political bias of media sources serves as a significant predictor of the likelihood of propagandistic

Table 3.8: Data distribution for the English and Spanish DIPROMATS corpora.

Class	**Train (ENG)**	**Test (ENG)**	**Train (SPA)**	**Test (SPA)**
Propaganda	1,974	N/A	1,199	N/A
Non-propaganda	6,434	N/A	4,921	N/A
Country				
China	2,170	852	2,178	819
European Union	2,043	873	1,508	957
Russia	2,005	955	795	596
USA	2,190	924	1,639	1,099
Type of tweet				
Tweet	6,742	2,856	3,586	2,302
Quoted	825	356	888	541
Retweet	473	227	1,221	401
Reply	368	165	425	227

content being used and that a smaller user community tends to disproportionately disseminate such articles. Furthermore, they found that forums focused on less mainstream parties in a country tend to share more biased news, and that cultural distinctions may influence the propaganda strategies used. Additionally, they noted that submissions or comments containing a higher volume of propaganda are likely to garner greater user interaction, either assessed through the quantity of comments or through upvotes and downvotes.

Table 3.9: Reddit dataset distribution, adapted from [46]

Subreddit	**Submissions**	**Comments**
Politics	317K	20M
Democrats	9.8K	54K
Republican	8.2K	41K
UKPolitics	42K	1.8M
LabourUK	7K	58K
Tories	1.1K	12K

3.2.3 Propaganda Disseminated in Facebook

Pushing for a new modality in detection of persuasion techniques in images and texts, the organizers of SemEval-2021 Task 6 [47] used a list of 22 techniques based on previous propaganda research (20 of them applicable to text and 2 to images) to label a collection of memes from Facebook. A total of 26 groups discussing themes such as politics, vaccines, and gender equality were crawled from 2020. The annotation step was executed in two phases: 1) Independent annotation of memes by annotators, and 2) Final gold labels by all annotators and a consolidator. Their final corpus consists of 950 memes, each meme containing at least one persuasion technique.

3.3 Discussion of Related Work Shortcomings

Figure 3.1 provides a summary of the various related works that are pertinent to this particular study focused on the field of computational propaganda detection. This summary not only highlights the significant contributions made by various researchers over time but also outlines the limitations that have been identified within these works. We aim to address these limitations through the efforts and findings of our own research.

When it comes to news articles, there are detection tasks aimed at document level and sentence level. The techniques used to detect propaganda on them mostly involve some kind of transformer-based classifier, either stand-alone or in an ensemble in addition to deep learning models. There exists some studies of propaganda on Twitter, with one using older pre-existing collections, and a Reddit study focused on political forums from the USA and the UK. However, by analyzing the related work, we identified some research opportunities further described in the following subsections.

3.3.1 Scarcity of Data and Format Differences

The first inclusion of propaganda in the TSHP-17 dataset shows an area of improvement in terms of considered number of propagandist sources, but also inconsistencies in number of documents. For example, although their creators claim to have over 74k articles, their publicly distributed files only account for approx. 39k articles. Barrón-Cedeño et al. elaborates on this matter, taking into account more propagandist sources but also describing a more realistic number of documents [31]. Yet, the number of resources aimed specifically towards propaganda detection on social media is still considerably low, not to mention the fact that texts from Twitter are by their nature noisy (they are brief, contain platform-specific features, and are riddled with typos and grammatical errors [48]).

3.3.2 Manual Annotation and Distant Supervision

As noted in a study of related matters about political ideologies [26], a carefully annotated corpus by experts may end up being relatively small, so the authors suggest that future work may explore semi-supervised models or active learning techniques for annotating and preparing larger corpora. Every classifier needs quality data to make good predictions. Annotation paradigms can be organized in supervised, unsupervised, and alternative approaches. As part of the latter, the distant supervision scheme, initially conceived for relation extraction purposes [49], relies on an external database to provide the labeled sources of information to subsequently create instances from them for training data. The labels produced by manual-annotation efforts by experts are considered of higher quality in comparison to distant supervision, however, this paradigm does not suffer from some of the disadvantages of hand-labeled efforts, such as being expensive, time consuming, and limited in quantity.

3.3.3 Contextual Information

There are different perspectives in the form of contextual information that can be further analyzed to unravel social patterns, and explored towards building a more complete solution to detect propaganda:

- *Bias levels* and *geographic origins*: Aside from "non-propagandist" or "propagandist" labels, more dimensions can be associated to news sources, such as their bias levels (from "Extreme-Left" to "Extreme-Right" ideologies) and their country as the place where the news feed is established.
- *Emotions*: Some of the most used propaganda techniques are associated with emotions, this suggests that they play an important role in manifestation of propaganda [10], [50].

3.3.4 Concept Drift

A prior study of computational propaganda used pre-existing Twitter datasets [10], however, an issue lies in the timing of the publication of said collections. Due to the dynamic nature of the news domain, the occurrence of concept drift leads to a stagnation of models trained on historical data, resulting in a decline in performance [51]. This element, along with manual annotation schemes, holds particular importance within the realm of social media under investigation in this research, which is susceptible to quick temporal changes in topics, the introduction of new terms with variable discriminative power, vanishing of classes and rise of modern fields [52].

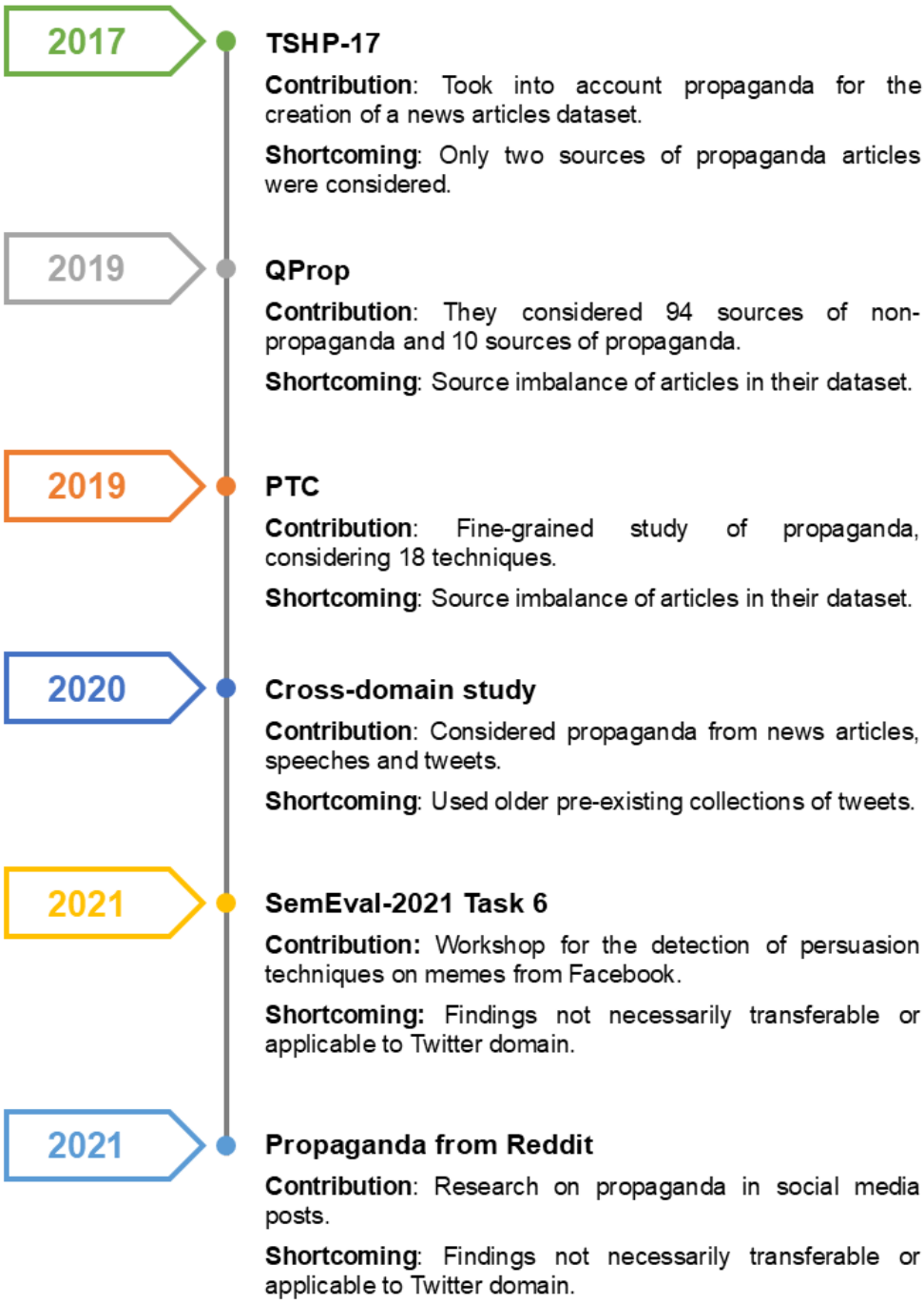

Figure 3.1: Contributions and shortcomings of relevant related work about propaganda detection.

Chapter 4

Propitter, a Corpus of Propaganda in Twitter

Here, we focus on investigating the propaganda spread on Twitter by media outlets that have been deemed unreliable or questionable due to their promotion of propaganda. Thus, one of our main contributions is the development of a new Twitter corpus for computational propaganda detection. This corpus, which we named as *Propitter*, to the best of our knowledge is the largest of its kind, containing more than 385 thousand tweets from more than 240 news sources accounts.

Propitter distinguishes itself from prior propaganda collections by being built by extracting information from numerous Twitter accounts associated with "everyday" news sources. They were chosen based on an external knowledge resource involving propaganda bias in news media. Its construction consists of data collected by distant supervision (Section 4.1.1), cross-domain filtering (Section 4.1.2), and in-domain data expansion (Section 4.1.3). These stages are further explained in the next sections.

4.1 Construction Methodology

4.1.1 Stage 1: Data collection by distant supervision

The construction of *Propitter* was inspired by *QProp* [31], which encompasses news articles from 10 propagandist and 122 non-propagandist sources (see further details in Section 3.1.2). First, we used Media Bias/Fact Check[1] to address the imbalance in the number of sources per class. The Media Bias/Fact Check website has assigned news sources in different *"Bias categories"*, where propagandist sources are found within *"Questionable sources"*. Each questionable source may or may not have the *propaganda* label as *"Questionable Reasoning"*. Using this criteria, we reviewed the news sources to identify those that tend to disseminate propaganda content and those that do not and then were labeled as *propagandist* or *non-propagandist*, respectively. For the sources having a Twitter account, we retrieved posted tweets using the Twitter API[2]. A total of $635\,k$ tweets published between January and August of 2021[3] were gathered from 244 distinct sources (the complete list of sources can be consulted in Appendix A1 and A2). Two filtering criteria were applied. A tweet was discarded if it was identified as being written in other languages than English[4], or if it contained at least three trending topics on the date of publication[5]. These heuristics aim at minimizing spam [53]. At the end of this stage, there were a total of $545,997$ tweets.

[1] https://mediabiasfactcheck.com/

[2] `https://developer.twitter.com/en/docs/twitter-api`

[3] *Propitter* includes a kind of bonus partition called *"Train ('17-'18)"*, with data collected from a similar time period as *QProp*, with the purpose of having in the future the opportunity to conduct cross-domain analyses.

[4] They were identified using Polyglot v. 16.7.4; `https://polyglot.readthedocs.io/en/latest/`.

[5] Using Trend Calendar, `https://us.trend-calendar.com/`

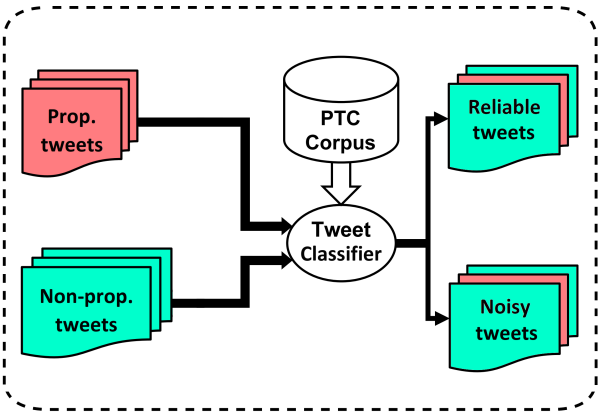

Figure 4.1: A classifier trained on the *PTC* corpus (white arrow) makes predictions over the *propagandist* (red) and *non-propagandist* (green) tweets collected in the previous stage. If both prediction and pseudo-label are the same (regardless of class) for a tweet, it is considered to be *reliable*, or *noisy* otherwise.

4.1.2 Stage 2: Cross-domain tweets filtering

As expected, the collected tweets are not free of noise due to their automatic labeling through distance supervision. Attempting to enhance the quality of the corpus and reduce the number of noisy tweets, we implemented the filtering process represented in Figure 4.1. This process capitalizes on the *PTC* corpus [28] (see Section 3.1.3 for details), which is manually labeled at the sentence level having a similar length in words to tweets (on average 23 and 20 tokens, respectively). *PTC* was used to fine-tune a base-uncased BERT model [22] for classifying the tweets gathered in the previous step as *propaganda* or *non-propaganda*. We adopted a similar methodology to that described in [34], however, upon replicating their system, superior performance was achieved through the use of a single BERT model (see more details in Section 3.1.3). Considering the classes provided through distance supervision as pseudo-labels for the tweets, we compare these classes against the ones obtained with the binary classifier. When both labels coincide (regardless of the class) the tweet is considered as *reliable*, otherwise as *noisy*.

A total of 337,155 tweets were judged as *reliable*, whereas 208,842 were flagged as *noisy*. Table 4.1 shows a few instances of the former, while Table 4.2 shows some instances of the latter. It is worth noting that, in Table 4.2, the first three tweets use propaganda techniques to convey their message: *exaggeration*, *name-calling*, and *appeal to fear*, verbalized by words such as *worst*, *savages*, and *dread*, respectively. The last four tweets include controversial topics (such as *price hikes*, the *Chinese political system*, and *COVID-19*), which are very likely discussion triggers. Besides, in these examples no propaganda techniques can be identified at first glance; facts are simply mentioned more objectively. Broadly speaking, what the classifier seems to do is filter out certain types of (presumably) false-negative and false-positive samples.

Table 4.1: Examples of *reliable* tweets at Stage 2. For these samples, the distant supervision (DS) pseudo-labels and the classifier's (CLF) predictions agreed on the class assignment.

Sample tweets	**DS label**	**Stage 2 CLF**
Palestine's Petty Fiefdoms: How the Palestinian Authority and Hamas are Destroying the Dream of a Free Palestine with Torture, Corruption and a Parallel Police State URL	Propaganda	Propaganda
After 30 Years of Brutal Rule, Sudan's Regime is Crumbling Under the Weight of a New Movement URL	Propaganda	Propaganda
#Olympics: British long jumper @USER says athletes understand its still a competition despite no fans #Tokyo2020 #TokyoOlympics URL	Non-propaganda	Non-propaganda
Duchess of Cambridge Kate self-isolating after COVID-19 contact URL URL	Non-propaganda	Non-propaganda

4.1.3 Stage 3: In-domain data expansion

In Stage 2, a set of *noisy* tweets was identified, possibly mislabeled based on the discrepancy between the initial distance supervision assessment and the binary

Table 4.2: Examples of *noisy* tweets filtered by the classifier at Stage 2. For these samples, the distant supervision (DS) pseudo-labels and the classifier's (CLF) class probabilities are at maximum disagreement.

Sample tweets	**DS label**	**Stage 2 CLF**
@USER Oh my God. Two of the world' worst mass murderers.	Non-propaganda	Propaganda
'Savages!': Ukraine's Black Olympian and Lawmaker Says He Was Verbally Attacked, Called 'Black Monkey' After He Won The Nation's Sole Tokyo Gold Medal URL	Non-propaganda	Propaganda
The Taliban's stunningly swift takeover of Afghanistan has caused dread across much of the nation, as Afghans anxiously readjust to life under a militant group that repressed millions when last in power. URL	Non-propaganda	Propaganda
Dollar exchange rates in Iraq URL	Propaganda	Non-propaganda
Gas prices expected to increase by up to 20 cents over the summer URL	Propaganda	Non-propaganda
The Communist Party of China and Kenya's Jubilee Party will take practical measures to further strengthen cooperation and exchanges. URL URL	Propaganda	Non-propaganda
UAE reports 1,321 new coronavirus cases, 3 deaths URL URL	Propaganda	Non-propaganda

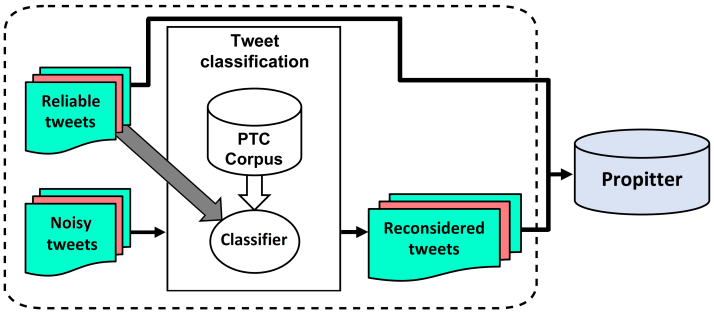

Figure 4.2: The same classifier from Stage 2 is re-trained with all *reliable* tweets (gray arrow) and then makes a second round of predictions over the *noisy* tweets. If the new prediction agrees with the pseudo-label for a tweet, the sample is reconsidered into *Propitter*. At the end of this stage, *Propitter* is formed by joining the sets of *reliable* and *reconsidered* tweets.

classifier. Nonetheless, this mismatch can potentially be attributed to the distinctive ways *propagandist* and *non-propagandist* contents are expressed in tweets and news articles. For example, in the case of Twitter, it is very common for posts to contain a short sentence accompanied by hashtags and URLs. Hence, to reconsider certain tweets that may have been misclassified as *noisy*, we conducted an expansion procedure inspired by [54]. First, the *reliable* tweets identified in the previous stage were merged with the instances of the *PTC* corpus. Then, using these data comprising tweets and news articles' sentences, we fine-tuned another BERT base-uncased model. Our intuition is that this second binary classifier is better aligned with the inherent characteristics of the language used in tweets. All the *noisy* tweets are then passed through a second classification round. The obtained labels are compared against the distance supervision pseudo-labels from Stage 1 once again and those that coincide are incorporated into the *Propitter* dataset. Figure 4.2 shows a schematic representation of the in-domain data expansion procedure. This stage starts with $208,842$ *noisy* tweets and, after the second classification, $48,236$ tweets were reconsidered as a second batch of reliable tweets.

Table 4.3 shows instances where the predictions made by the first classifier

Table 4.3: Examples of *reconsidered* tweets filtered by the classifier of Stage 3.

Sample tweets	**Stage 2 CLF**	**DS label & Stage 3 CLF**
@USER @USER How much is the CCP paying you to spread disinformation?	Non-propaganda	Propaganda
Here are "10 Reasons Why Abortion Is Evil" and must be opposed. We encourage you to share this. URL #prolife #abortion #tfp #catholic	Non-propaganda	Propaganda
With video. Of course they lied. They're Democrats... Jason Chaffetz says it appears they even violated House rules on using deceptive video. URL #tcot #MAGA #ImpeachmentTrial	Non-propaganda	Propaganda
Colorful toy fish, bouncy balls, and stuffed animals are just some of the surprises frozen into whimsical sculptures made by an Osaka icemaker. URL URL	Propaganda	Non-propaganda
The hoodie is either a company mistake or a reference to former USC President Robert Caslen's 2021 graduation speech where he mistakenly congratulated alumni of the 'University of California.' URL	Propaganda	Non-propaganda
China is facing a high-profile test of its commitment to curbing industrial pollution after steel output has surged to well beyond its target of capping production at 2020's peak. URL	Propaganda	Non-propaganda
France accuses Erdogan of 'provocation' over Cyprus visit, remarks URL	Propaganda	Non-propaganda

(which was initially trained solely on sentences extracted from news articles) changed upon being trained with tweets. These tweets, in comparison to those in Table 4.2, exhibit distinctive features that are specific to Twitter, such as references to users, hashtags, and URLs. After being exposed to tweets containing propaganda the classifier could recognize patterns associated with these attributes of the Twitter domain. The first three tweets employ propaganda techniques such as *reductio ad hitlerum*, *name-calling*, *slogans*, and *loaded language*. On the other hand, the last four tweets do not exhibit any discernible technique; instead, they directly discuss certain events and appear to be grounded on factual information.

Table 4.4: Examples of discarded or *non-reconsidered* tweets from Stage 3.

Sample tweets	**Stage 2 & 3 CLF**	**DS label**
Big things are happening in one of the world's smallest capitals. The Arctic city of Nuuk in Greenland is poised to become the world's first certified "sustainable capital" by the Global Sustainable Tourism Council URL	Propaganda	Non-propaganda
The Taliban promises women's rights and security under Islamic rule, but many Afghans are desperate to flee. URL	Propaganda	Non-propaganda
Catch me live on The Sons of Liberty radio show Now! URL	Non-propaganda	Propaganda
UK must welcome 'tens of thousands of Afghan refugees', urges Labour URL	Non-propaganda	Propaganda

Table 4.5: General statistics of *Propitter*, showing the total number of tweets in each partition, as well as the portion corresponding to *propaganda* tweets. The "Bonus" training partition refers to a small subset of data with timestamps similar to *QProp*, created to enable and promote cross-domain experiments between collections.

	Partition	**Tweets**	**Propaganda**
Main	Train	293,480	77,167
	Development	38,511	6,454
	Test	38,501	7,045
Bonus	Train ('17-'18)	14,899	10,509
Total		385,391	101,175

Table 4.4 shows examples of tweets that were discarded, or in other words "non-reconsidered", because even the predictions made by the Stage 3 classifier were not in agreement with the pseudo-labels assigned by distant supervision.

After the data collection, filtering, and expansion stages, the number of tweets in *Propitter* was depurated from $635\,k$ to $385\,k$ instances. In order to establish partitions for training, development, and testing, the tweets were arranged in chronological order. The training set consists of the 80% of the earliest tweets followed by 10% for validation, and the remaining 10% for testing. The final data distribution of *Propitter* is shown in Table 4.5.

Table 4.6: Classification baseline results on *Propitter*. All experiments were carried out on its "Main" partitions, reporting measures over the *propaganda* class.

Classifier	**Precision**	**Recall**	F_1**-score**
BERTweet	78.46 ± 2.73	85.43 ± 2.38	81.72 ± 0.43
Linear-SVM	68.05	72.68	70.29

4.2 Propitter's Classification Results

Addressing propaganda detection as a binary classification problem (*propaganda vs. non-propaganda*) is the main task that can be performed using *Propitter*. In this sense, two baselines are proposed: a Bag-of-Words (BoW) representation with a Linear SVM [55], and BERTweet [56], a pre-trained transformer-based model. The obtained results are shown in Table 4.6. Here, it is important to highlight that all the experiments hereafter were performed using BERTweet for three main reasons: a) this kind of classifier performs very well in a wide variety of NLP tasks, b) it is a pre-trained model on Twitter data, and c) it shows a better performance compared to a BoW approach in the baseline results. The model parameters (which were chosen after multiple tuning iterations) are a batch size of 32, a learning rate of $2e-5$, an Adam optimizer, and 3 epochs. We report the average of running it five times since the fine-tuning process of BERTweet is not deterministic.

4.3 Propitter's Qualitative Analysis

With the intention of providing additional information about *Propitter*, we carried out a qualitative analysis of it. Inspired by the analysis of prominent linguistic attributes in propaganda phenomenon on the *TSHP-17* dataset presented in [29], we calculated the same set of features on *Propitter*: the number of second-person pronouns, superlative adjectives, and weak subjective words (those that might

Table 4.7: Linguistic features and their average occurrence ratio between *propagandist* and *non-propagandist* tweets (second column) and news articles (third column). A ratio above 1 means the feature is more frequent in the *propaganda* class. The values from the third column are borrowed from [29].

	Ratio	
Lexicon markers	***Propitter***	***TSHP-17***
2nd person (You)	1.87	6.73
Weak subjective words	1.40	1.13
Superlatives	1.00	1.17

only have particular subjective uses, according to [57]). The obtained results are shown in Table 4.7, with the relative frequencies of second-person pronouns and weak subjective words indicating that these lexicon markers are more associated with propaganda in both articles and tweets.

As previously mentioned, propaganda involves the use of different techniques to achieve its purposes. Given that *Propitter* has been developed with a hybrid approach for data labeling, it is interesting to observe whether or not such a method allows us to include samples using any propaganda technique. Table 4.8 presents some propagandistic examples manually identified where a particular technique was used. However, it is important to mention that identifying propaganda techniques in tweets is beyond the scope of *Propitter* and that this information is included only for illustrative purposes.

4.4 Creating *PropitterX*: Adding Contextual Information

Propaganda manifests itself in various forms [28], which can be associated to factors such as political biases [58] and emotions [3]. Intending to contribute to the study of propaganda from different perspectives, we extend *Propitter* by incorporating four kinds of contextual features:

Table 4.8: Sample tweets from *Propitter* that display the use of different propaganda techniques in the collection.

Sample tweets	**Propaganda techniques***
Egyptians across the political spectrum are **outraged** by a "politicized" European Parliament resolution that they call a blatant intervention in Egypt's internal affairs, which serves the interests of **terrorists fighting** the government of President el-Sisi. URL	Loaded language
AVERAGE JOE? Biden Says Far-Left 'Doesn't Like Him' Because He Blocks Their 'Socialist Agenda' URL	Name calling/ labeling
Erdogan Terrorists Open Fire at Other **Erdogan Terrorists** in Al-Bab City - Video: URL #Syria #News #Politics #Quneitra #alBab #Aleppo #Turkey **#Terrorism #Erdogan** #alQaeda #FSA #Nusra #ISIS #HTS #NATO #RegimeChange #USA #Russia	Repetition
Raise your hand if you thoroughly enjoyed watching Trump lose the election for the **millionth time today**! (Raised Back of Hand Emoji Raised Back of Hand Emoji Raised Back of Hand Emoji)	Exaggeration or minimization
Donald Trump and his acolytes say poor white Americans are victims, **but are they?** URL	Doubt
Sick. Kristol that is. As for the center they are simply information deprived. **How does opening borders in the midst of a global pandemic, and inviting refugees from terrorist states, and removing the Houthi terrorists from terrorist lists help anyone but our enemies?**	Appeal to fear/prejudice
In dealing with the COVID crisis, the public health experts **have failed the nation**, betrayed their mission and spread confusion. So many outrages have been committed in the name of "science" that people are rightfully distrustful. URL #tfp #Covid	Flag-waving
The (Second) Horror of the Flint Water Crisis: By Walter Block - **If** the drinking water from the Flint River in Michigan looked dirty, or smelled bad, the disaster could probably have been avoided. **No one would have drunk the poisonous liquid**, and roughly... URL	Causal oversimplification
"AND WE WILL **MAKE AMERICA GREAT AGAIN**!" WOW! THANK YOU PRESIDENT TRUMP!!! URL	Slogans

* *Propitter* has binary labels rather than multiclass technique labels.

Political bias Media Bias/Fact Check subjects information sources to a process of analysis of the news and opinions they publish[6]. This platform assigns a "Bias Rating" to each source according to the political perspective they promote in a variety of categories including general philosophy, economic policies, and education, among others[7]. Based on this, we assign each instance in *PropitterX* with its corresponding political bias, based on the labels included in Appendix A3, which range from extreme-left to extreme-right. More details of the distribution of grouped biases are also shown in Table 4.9.

Table 4.9: Bias statistics of main partitions from *PropitterX* corpus

Bias	**Train**	**Dev**	**Test**
Left-wing	142,686	19,270	16,999
Right-wing	119,823	13,619	14,805
Misc.	30,971	5,622	6,697

Temporal split Each instance from Twitter, nowadays known as X, has a set of metadata including the date and time of its publication. This information allowed to chronologically organize the collection, and subsequently add a temporal split attribute to the posts in the main *Train* partition.

Affective information We have incorporated an additional feature which indicates the main emotion evoked by each tweet (more details in Table 4.10). Specifically, we have taken into account Ekman's categorical model of emotions [59] which considers: *fear*, *anger*, *joy*, *sadness*, *surprise*, *disgust*, and *neutral*. To assign the most likely emotional category related to each tweet,

[6]While it is difficult to verify the veracity or accuracy of this platform with respect to the labels it assigns to news sources, there is a precedent of having been used in the related work consulted for the creation of a corpus of propaganda articles [31]. In addition, Media Bias/Fact Check provides details on the methodology and rating system they use in their own media analysis, including their own references, available at `https://mediabiasfactcheck.com/methodology/`.

[7]`https://mediabiasfactcheck.com/left-vs-right-bias-how-we-rate-the-bias-of-media-sources/`

we employed a BERT model fine-tuned[8] [9] with a Twitter sentiment analysis dataset[10].

Table 4.10: Emotion statistics of main partitions from *PropitterX* corpus

Emotion	Train	Dev	Test
Anger	47,436	5,792	5,305
Disgust	2,279	280	217
Fear	148,123	20,436	21,885
Joy	20,727	2,430	2,162
Neutral	61,400	7,766	7,194
Sadness	10,127	1,446	1,336
Surprise	3,388	361	402

Geographic origin The vast majority of sources that were referenced have their country of origin available as informative data in the MediaBias/Fact Check web resource. We compiled this information and, motivated by how data is structured in related research [45], [46], we categorized the list of countries into five distinct regions: *America*, *Asia*, *Europe*, *Middle East*, and *Others*. Then, we associated each instance in *PropitterX* with one of these general geographic regions (more details are shown in Table 4.11).

Table 4.11: Region statistics of main partitions from *PropitterX* corpus

Region	Train	Dev	Test
America	201,210	22,755	22,433
Asia	26,380	5,991	6,544
Europe	18,947	4,084	3,908
Middle East	26,970	4,201	4,225
Others	19,973	1,480	1,391

[8] `https://huggingface.co/bhadresh-savani/bert-base-uncased-emotion`
[9] Achieving a 0.94 in F1-score on Emotion Dataset from Twitter.
[10] `https://huggingface.co/datasets/philschmid/emotion`

4.5 Summary

Propaganda is pernicious and takes advantages of the widespread use of social networks. Our research addresses the challenge of identifying propaganda on Twitter by employing a construction process that leverages on pre-existing resources from the news article domain to clean a collection of data gathered under a distant supervision scheme. This allowed us to create a corpus in which we evaluated classification approaches as baselines. As a consequence, we discovered that a state-of-the-art transformer-based classifier is, as expected, more resilient than other alternatives (such as SVM with Bags-of-Words) in terms of being less affected by variables like temporal placement and volume of training data. Considering these two variables, we observed that the chronological order of data affects more than the amount of data (volume) while training a classifier. We expect our corpus to be a useful resource in the area of computational propaganda detection, since it would be the first collection of tweets in English specifically built for this task.

Chapter 5

The Influence of Contextual Features for Propaganda Detection in Tweets

In order to assess the role and relevance of the contextual attributes for the identification of propagandist tweets, we propose four experimental scenarios described in the following subsections. All of these scenarios are based on a binary classification task, but each considering a different subset of tweets organized according to each of the contextual attributes. This allows us to shed light on their particular relevance in the study of propaganda on Twitter.

5.1 *PropitterX-LR*: On the Role of Political Bias

According to [60], there are grounds to suspect that political affiliation plays an important role in determining peoples' perception of reliable or unreliable sources of information. In this sense, we explore if this observation can be useful to find out if the propaganda produced by one political position differs from that of another. The inclusion of political bias information in *PropitterX* opens the

door to address these inquiries within the Twitter domain. For this purpose, we propose a subset of this collection referred to as *"PropitterX-LR"* to denote the "Left vs Right" case study. We consider the main *Train*, *Development*, and *Test* partitions of *Propitter* and divide each partition into two groups: *Left Wing* (encompassing Extreme Left, Far-Left, Left, and Left-Center) and *Right Wing* (Extreme Right, Far-Right, Right, and Right-Center) according to the political bias associated to each instance, resulting in the partitions shown in Table 5.1.

Table 5.1: Statistics of *PropitterX-LR* according to the amount of left-wing and right-wing tweets per partition.

Bias	**Train**		**Dev**		**Test**	
	Prop.	Non-prop.	Prop.	Non-prop.	Prop.	Non-prop.
Left Wing	10,831	131,855	963	18,307	885	16,114
Right Wing	62,952	55,737	5,047	8,534	5,680	9,106

For experimental purposes, two binary *propaganda* vs *non-propaganda* classifiers were designed: a *"Left Wing Classifier"* (LWC), trained and validated using solely left-biased data, and a *"Right Wing Classifier"* (RWC), trained solely on right-biased data. Both classifiers were evaluated on test data with the same and different political biases. Table 5.2 shows the obtained results, where the average classification rate together with its corresponding standard deviation value are presented. Both classifiers struggle to identify propaganda from the opposing side in the political spectrum properly. As a result of the data distribution within the partitions in Table 5.1, RWC tends to classify a greater number of tweets as *propaganda*. Given that the dominant class in the left-wing part of the test set is *non-propaganda*, the precision of RWC is low while its recall is high. Conversely, LWC is inclined to classify more tweets as *non-propaganda*. Since the prevalent class in the right-wing part of the test set is also *non-propaganda*, the precision of LWC is high while its recall is low.

A study was conducted to further comprehend the rationale behind this difference in performance through an analysis of the topics addressed by each

Table 5.2: Results of the political bias experiment. The evaluation measures were calculated over the propaganda class.

Test Data	Classifier	Precision	Recall	F_1-score
Left Wing	LWC	81.59 ± 6.39	77.15 ± 6.91	78.82 ± 2.50
	RWC	26.92 ± 4.47	76.84 ± 6.85	39.39 ± 3.96
Right Wing	LWC	88.88 ± 3.10	40.81 ± 7.93	55.31 ± 7.29
	RWC	77.27 ± 2.05	91.53 ± 2.30	83.75 ± 0.85

group of biases. Employing *Latent Dirichlet Allocation* (LDA) [61], we analyzed the propaganda tweets corresponding to each political bias, and for each case, we identified its top 5 topics and their corresponding 10 most relevant words. Figure 5.1 shows that left-wing propaganda primarily emphasizes the pandemic and international matters in discussions concerning world-specific regions, such as the *US* and *China*. Conversely, right-wing propaganda places greater emphasis on topics involving *Joe Biden*, *abortion*, and *racism*. Overall, the word clouds below illustrate the relevance of political bias for automatic propaganda detection.

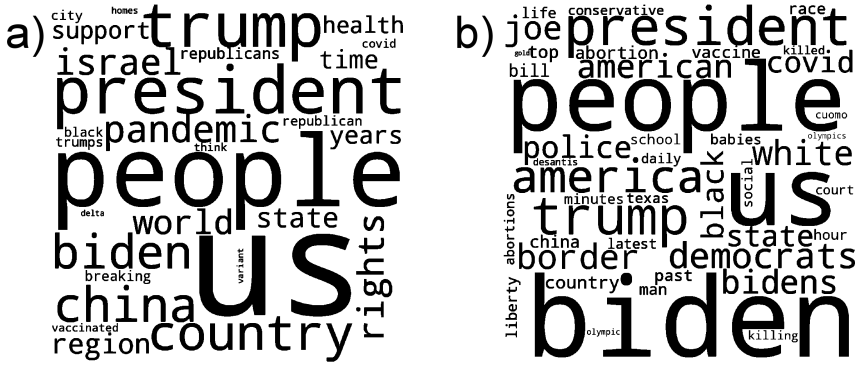

Figure 5.1: Word clouds with most prominent words from the top 5 topics detected by LDA in (a) left-wing propaganda and (b) right-wing propaganda.

5.2 *PropitterX-TIME*: On the Evolution of Trending Topics

The temporal evolution of a given phenomenon is a key challenge when addressing a problem as a classification task under a supervised approach [62]. In the case of propaganda detection, changes like the emergence and disappearance of topics of public interest, as well as the use of terms associated with a particular intention could have a significant impact on the performance of the classifiers [63]. In this sense, it is imperative to explore the extent at which these changes harm the automatic detection of propaganda.

As previously stated, the tweets in *PropitterX* correspond to a period of six months and are ordered chronologically, with the training set having the oldest tweets and the test set having the most recent ones. With the intention of assessing the role of topic changes across time in propaganda, we propose a data arrangement called *"PropitterX-TIME"*. We split the training set into five partitions of 60 k tweets each, respecting the chronological order. The date ranges for each temporal split are shown in Table 5.3.

Table 5.3: Date ranges for each temporal split in PropitterX-TIME.

Split number	From	To
Split #1	1-Jan-2021	8-May-2021
Split #2	8-May-2021	18-Jun-2021
Split #3	18-Jun-2021	11-Jul-2021
Split #4	11-Jul-2021	27-Jul-2021
Split #5	27-Jul-2021	7-Aug-2021
Test Set	13-Aug-2021	20-Aug-2021

The idea is to train a separate classifier for each split (10% of each partition is used for validation) and compare the performance of the five resulting classifiers after generating predictions on the whole test set. As shown in Figure 5.2, the performance of the classifiers in detecting propaganda gradually improves as the

time span of the training data gets closer to the one of the test set.

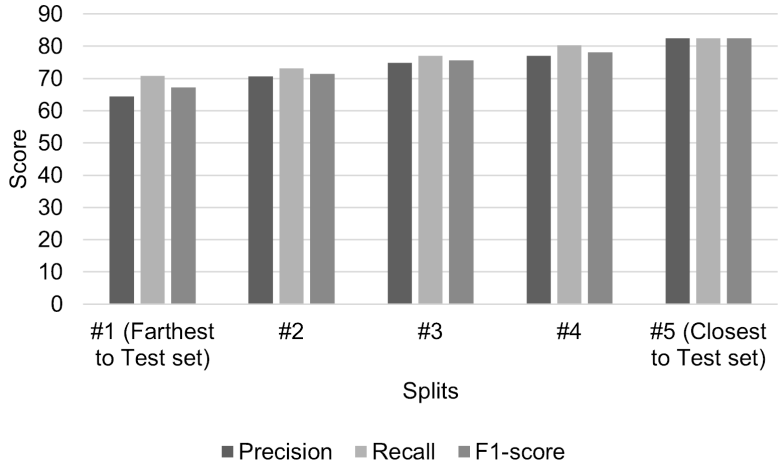

Figure 5.2: Classification results over the propaganda class with chronological training splits.

To enhance our understanding of the previous results, Figure 5.3 outlines the evolution of the topics across the different training splits, as well as their comparison with the main topics from the test set. For example, the term *trump* holds significance in splits #1 and #2, yet its importance gradually diminishes from split #3 onwards to the test set. The terms *covid* and *vaccine* start to emerge from split #3, with their frequency steadily increasing up to split #5. Terms like *military*, *taliban*, and *afghanistan* gain popularity starting at split #4, before turning the most crucial ones in the test set, temporarily overshadowing other subjects such as *covid*. In general, it is observed that the topics addressed in the test data are more related to the ones in the most recent training partitions.

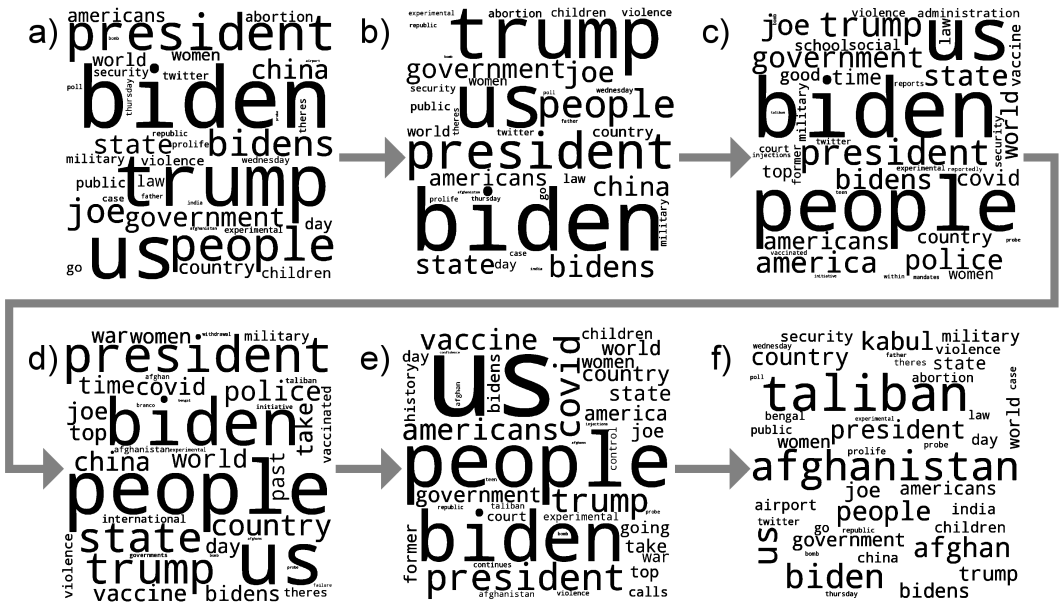

Figure 5.3: Word clouds with most prominent words in the top 5 topics detected by LDA in the propaganda from chronological splits (a) #1, (b) #2, (c) #3, (d) #4, (e) #5, and (f) Test set.

5.3 *PropitterX-EMO*: On the Relevance of Affective Information

Some widely recognized propaganda strategies aim to elicit an emotional response. This can be observed, for instance, in the use of *loaded language* and *slogans*, where words or expressions with emotional connotations are used to sway the audience's opinion [7], [28]. Thus, exploiting the role of emotions to identify propaganda is a subject that deserves to be investigated. Accordingly, we explore if a classifier trained with messages that evoke emotions performs better in detecting propaganda compared to one trained with neutral messages. To address this, it is possible to employ a sub-collection denominated *"PropitterX-EMO"*, relying on the affective information attributes delineated at the beginning of Section 4.4. In particular, we applied a pre-trained BERT model[1] to obtain predictions of

[1] `https://huggingface.co/bhadresh-savani/bert-base-uncased-emotion`

Table 5.4: Distribution of tweets per primary emotion evoked and class in *PropitterX*. The percentage to the right of each emotion in the first column corresponds to its rate of occurrence in the training set.

	Train		**Dev**		**Test**	
Emotion	Propaganda	Non-propaganda	Propaganda	Non-propaganda	Propaganda	Non-propaganda
Fear(64%)	35,091	113,032	3,086	17,350	3,777	18,108
Anger(20%)	21,169	26,267	1,570	4,222	1,438	3,867
Joy(9%)	4,111	16,616	360	2,070	360	1,802
Sadness(4%)	2,399	7,728	217	1,229	257	1,079
Surprise(2%)	1,594	1,794	131	230	175	227
Disgust(1%)	1,387	892	134	146	117	100
Neutral	11,416	49,984	956	6,810	921	6,273

the emotions associated with each message. Table 5.4 shows the distribution of tweets in terms of emotional categories automatically determined.

In order to evaluate the relevance of emotional information in propaganda detection, we designed two training settings using *PropitterX-EMO*: *a*) Considering only tweets regarding any of the 6 emotional categories taken into account (i.e., *fear*, *anger*, *joy*, *sadness*, *surprise*, or *disgust*), and *b*) Considering only *neutral* tweets, i.e. those in which no salient emotion was found. A total of $60\,k$ neutral tweets were sampled to train the *neutral classifier*. Then, to have similar training conditions, we matched that same training volume for the *emotional classifier*, taking into account the proportion of occurrence of each emotion in the training set (both classifiers are referred to in this way, *neutral* and *emotional*, hereinafter). Further details can be consulted in Appendix A4.

The *neutral classifier* was trained using 48,853 *non-propaganda* tweets and 11,147 *propaganda* tweets. The *emotional classifier*, on the other hand, was trained with 30,000 *non-propaganda* tweets and 30,000 *propaganda* tweets. As depicted in Table 5.5, each classifier demonstrates a slightly better performance when the training and test conditions align. It is worth to highlight that despite the *neutral classifier* being trained with only a third of the volume of *propaganda*

instances compared to the *emotional classifier*, the performance gap between them is not too wide.

Table 5.5: Comparison of the performance of classifiers trained with: i) tweets that evoke a predominant emotion, and ii) neutral tweets; evaluation measures correspond to the propaganda class of the test set.

Test Data	**Classifier**	**Precision**	**Recall**	F_1**-score**
Emotional	*Emotional*	68.02 ± 4.52	91.01 ± 2.65	77.68 ± 2.30
	Neutral	82.55 ± 4.30	69.21 ± 4.47	75.04 ± 0.75
Neutral	*Emotional*	68.75 ± 7.94	88.12 ± 3.98	76.75 ± 4.39
	Neutral	78.91 ± 2.16	82.47 ± 2.98	80.61 ± 1.74

Figure 5.4 illustrates the themes deliberated in the propaganda of the top two prevalent emotions in *PropitterX*, namely *fear* and *anger*, in comparison to propaganda devoid of a dominant emotion. *Fear-based propaganda* delves into matters concerning warfare and public health (*pandemic*, *vaccine*, and *covid*). *Anger-provoking propaganda* tackles subjects associated with *racism* and *migration*. Lastly, while propaganda in *neutral* tweets shares certain terms with the aforementioned emotions, it presents a more uniform discourse where no particular topics emerge prominently (excluding discussions about D. Trump and J. Biden in the three cases). These findings, together with the results in Table 5.5, suggest that *neutral propaganda* contents cover a wide spectrum of subjects rather than focusing on a particular issue or trigger topic, and, therefore, that it can be a good starting point for training a general propaganda classifier.

5.4 *PropitterX-GEO*: On the Role of Region-Centered Content

Findings suggest that the nature of propaganda may vary depending on where it is generated [46]. This hypothesis seems plausible since a prominent technique employed in propaganda, denoted as *"flag-waving"*, involves exploiting deep-rooted

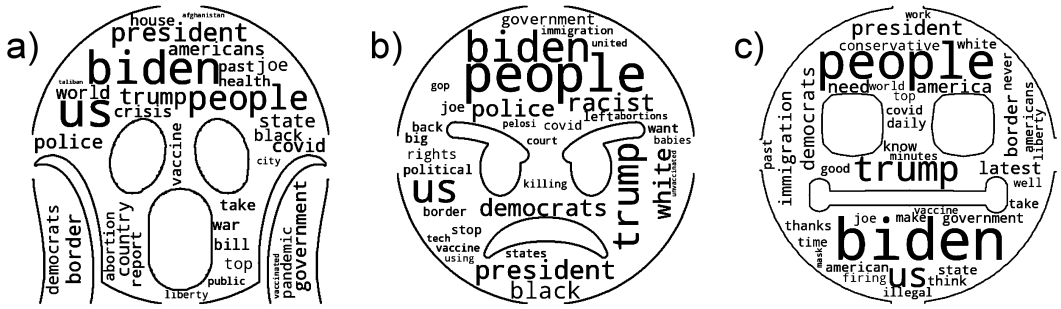

Figure 5.4: Word clouds with most prominent words in the top-5 topics detected by LDA in propagandist tweets that exhibit a predominant emotion of (a) fear, (b) anger, and (c) neutral tweets (no salient emotion found).

emotions of patriotism and nationalism to justify certain ideas or actions [28]. Taking into account this particularity a question arises, *Can propaganda from a particular region of the world be effectively used to identify propagandistic posts produced in a different geographic location?* Attempting to address this question, the *"PropitterX-GEO"* subcollection is proposed. It considers the source region attribute to split the data into five distinct groups: *America*,[2] *Asia*, *Middle East*, *Europe*, and a miscellaneous category known as *Others*, comprising Africa, Australia, and sources whose origin is unknown. The distribution of tweets per region in *PropitterX-GEO* is detailed in Table 5.6. It is worth mentioning that, unlike in the previous sub-collections where the *train* and *test* partitions were used as indicated in *Propitter*, in this case, all data were merged and then split according to the corresponding region of origin.

Table 5.6: Distribution of tweets per region in the *PropitterX-GEO* subcollection.

Region	**Propaganda**	**Non-propaganda**	**Total tweets**
America	64,555	181,843	246,398
Asia	5,448	33,467	38,915
Middle East	8,853	26,543	35,396
Europe	5,116	21,823	26,939
Others	6,694	16,150	22,844
		Total	370,492

[2]By America we refer to the continent.

Table 5.7: Results of training with one region and making predictions on the rest. The rows represent the regions used for training while the columns those used for testing. The evaluation measure is F_1 over the positive (propaganda) class. The best score per column appears boldfaced. There are no results on the main diagonal, since in each experiment all available tweets from a region were used to train the corresponding classifier.

Train Region	**Test Region**				
	America	Asia	Middle East	Europe	Others
America	–	38.24±5.51	**56.70±2.72**	47.07±3.64	**81.53±0.68**
Asia	68.33±2.75	–	54.34±6.01	21.13±3.82	55.93±4.49
Middle East	76.34±1.51	**48.66±3.24**	–	46.16±1.93	68.93±3.49
Europe	70.11±1.12	20.44±1.99	35.77±3.77	–	62.23±2.49
Others	**79.95±1.27**	32.45±3.04	50.74±2.49	**54.53±3.01**	–

Once the tweets were organized according to their corresponding region, we performed some binary-classification experiments to assess whether a classification model trained using examples of propaganda from a particular region of the world effectively identifies and differentiates propaganda from a different geographic location. Therefore, a classifier was trained with the data available from a particular region and evaluated over another one. The obtained results are shown in Table 5.7.

Training on data from *America* yields the most favorable detection results for propaganda from *Middle East* and *Others*. For *Asia*, it was better to train with data from the *Middle East*, and for *Europe* training with data from *Others* performed the best. Due to the comparable performance exhibited by classifiers trained with data from *America* and *Others*, we suspect that it is likely that the unspecified sources within the category *Others* are also located in America. On average, *Asia* is the region where it was most difficult to detect propaganda using off-region training data. It is important to emphasize that some of the propaganda techniques refer to aspects specific to the places where they are intended to be applied, such as *flag-waving* and *slogans*. Therefore, it is not surprising that there are differences between the propaganda spread in different regions.

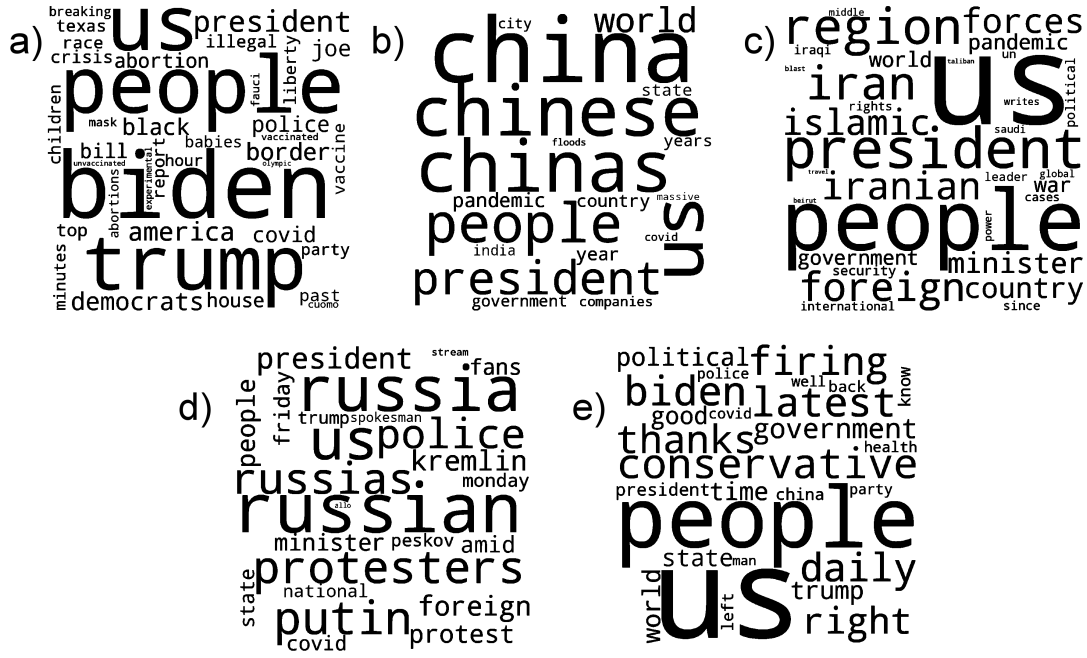

Figure 5.5: Word clouds with most prominent words in the top 5 topics detected by LDA in the propaganda from (a) *America*, (b) *Asia*, (c) *Middle East*, (d) *Europe*, and (e) *Others*.

As Figure 5.5 shows it is evidently discerned that the propaganda in each region references distinct entities and topics. The regions in which certain common propaganda terms are identified are "*America*" and "*Others*" (with a high frequency of references to D. Trump and J. Biden). The analysis of this figure, along with the results from Table 5.7 suggest that the classifiers are linking propaganda to region-specific tokens. We hypothesize that the amount of data used for training purposes did not make a significant impact on the outcomes (otherwise the classifier trained with data from *America* would have performed better in all experiments but this did not occur); rather, the key factor rested on the diversity in the topics represented.

5.5 Summary

In the previous chapter, we introduced *PropitterX*, a new resource developed on the basis of *Propitter*, a dataset of propaganda on Twitter. *PropitterX* includes tweets annotated for propaganda and incorporates various contextual information aspects such as political bias, geographical origin, emotions evoked in the messages, and temporal splits. These dimensions allow us to create some subcollections or data arrangements, amplifying the potential for conducting experiments that integrate propaganda with additional factors.

Some interesting insights into the association between propaganda and various of these contextual aspects were found through initial experimentation. The findings suggest that: *i*) Propaganda produced by sources with a left bias differs from that produced by a right bias; *ii*) Trending topics associated with propaganda seem to evolve, impacting the performance of the capabilities for recognizing propaganda: older messages are harder to identify than the most recent ones; *iii*) Neutral propaganda content covers a broader spectrum of topics than propaganda anchored in particular emotions; and *iv*) There is variability in propaganda across different geographical regions.

Chapter 6

A Contextual-Aware Approach to Improve Propaganda Classification

In the task of propaganda detection, some works have investigated the incorporation of context. The study conducted by [64] examined emotions and sentiments as means of communication and social influence. These characteristics were extracted using external models applied to tweets. For sentiments, the subcategories included *Positive*, *Negative*, and *Neutral*. For emotions the categories were *Anger*, *Joy*, *Optimism*, and *Sadness*. They "augmented" the original messages with textual features, such as *"The statement expresses optimism as emotional content. Its sentiment is positive. The message received 16 interactions. The country of origin is Russia."*. Janez et al. [65] explored the posting trends observed across different nations, such as patterns where countries that reference their own-origin within the content of their tweets tend to exhibit a higher likelihood of propagandistic behavior. They "injected" this information as context to their classification model by adding the phrase *"This has been written from [country]."* at the very beginning of each tweet, so that the model could consider the geographical context of each message during its classification process. Similarly, in our previous

research [66], we used the features of country of origin, tweet type, and emotion (determined using external models) as contextual keywords to leverage the classifier's architecture. This same approach explained in more detail in the section below, is examined in a more comprehensive manner in this study. Specifically, we assess the efficacy of the methodology in two distinct collections of tweets associated with both news outlets and government entities, with the latter collection providing an opportunity for a direct comparison of our findings with other classification systems [12]. In addition, we explore the *political bias* of the sources as an extra contextual feature, and alter the volume of training data to assess the relevance of contextual information in different classification scenarios. Finally, we also consider the automatic prediction of contextual attributes to evaluate the suitability of the proposed approach to situations in which this information is initially unavailable.

6.1 Contextual-aware Approach

Our proposed method is based on Bidirectional Encoder Representations from Transformers (BERT) models [22]. In BERT, a "sentence" refers to an arbitrary span of adjacent text, and a "sequence" indicates the input token sequence. The BERT model's input representation is designed to accommodate not just a singular text sentence but also a combination of two text sequences encapsulated in a single token sequence, where the initial token "[CLS]" holds the classification embedding and another special token, "[SEP]", is employed to separate segments or indicate the conclusion of the sequence [67].

Motivated by the incorporation of this auxiliary input in various other studies [67]–[69], we chose to leverage it as a mean to "provide context" to the sentence presented to the main input of text for the task of detecting propaganda. Figure 6.1 shows the method we employed to merge a tweet's content with its set of

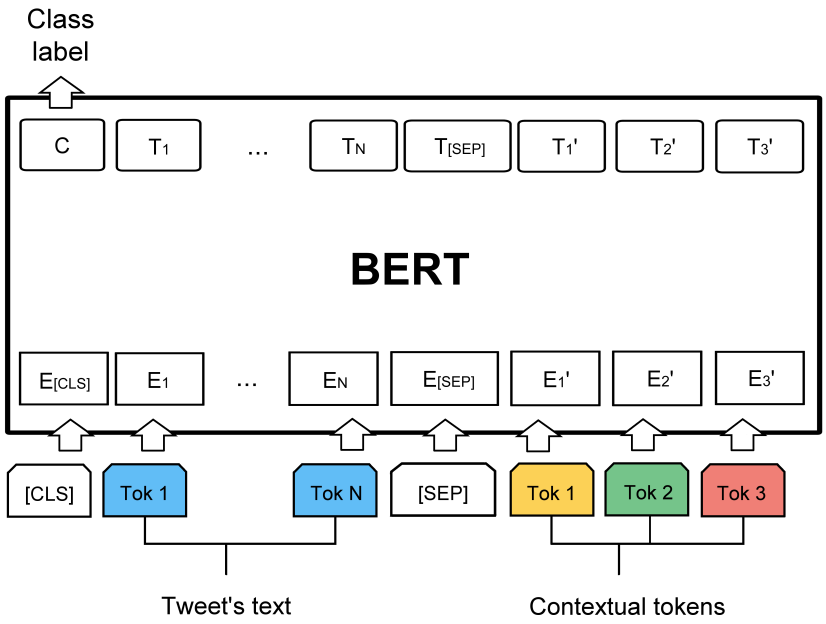

Figure 6.1: BERT's auxiliary input diagram with the contextual features concatenated to the tweet's text (adapted from [22]).

contextual features.

In this manner, we will assess the effectiveness of **BERT-CA**, a BERT model designed to be *context-aware*, which, in addition to using the tweet's text in the main input, will also be enriched with contextual attributes represented as "tokens" after the first [SEP] token in the secondary input. **BERT-CA** can incorporate various contextual tokens, and throughout our experiments we examined different types of context as well as combinations of these features to assess their effectiveness.

Table 6.1 offers some examples of sequence compositions that include context in the auxiliary input. In the first instance, the model is provided with the background that the tweet originates from a *left-wing American* perspective and that the content incites *fear* by questioning the safety of vaccine administration, particularly for pregnant women (it is important to note that the timeframe when this data was collected is related to COVID-19). In the second instance,

Table 6.1: Examples of input token sequences.

Input Sequence
[CLS] What are the vaccine risks for pregnant women? UW doctors have a few thoughts. URL [SEP] LEFT AMERICA FEAR [SEP]
[CLS] #Crimea is a vital part of Russian civilization. It is the point of origin of Russian Christianity, was in Ancient Rus and Russian Empire, Soviet Russia and USSR, reunited with Russia - dear to the hearts of all Russians URL [SEP] RUSSIA [SEP]
[CLS] Today marks a major milestone in making Europe the first climate neutral continent in the world. With the new target to cut EU greenhouse gas emissions by at least 55% by 2030, we will lead the way to a cleaner planet and a green recovery. [SEP] JOY [SEP]

the tweet contains strong nationalist sentiments towards Russian territory, which becomes clearer when acknowledging that this tweet originated from Russia, indicating they are expressing favorable views about themselves. The third instance discusses a significant achievement for Europe in the realm of climate change, advocating for a decrease in gas emissions towards a cleaner planet. The main emotion identified in this tweet was *joy*, which aids the model in recognizing that it serves as propaganda infused with emotional appeal.

6.2 Experimental settings

6.2.1 Dataset

To carry out our experiments, we used the *PropitterX* dataset (described in Chapter 4). It comprises a compilation of tweets originally posted by more than 240 prominent news media accounts, labeled in a binary manner as *propaganda* and *non-propaganda*. The dataset is structured into three primary sets: *Train*, *Dev*, and *Test*. In addition to the binary label of *propaganda* or *non-propaganda*, the dataset offers the following three contextual features per tweet:

- **Bias**: Political affiliation of the account responsible for posting the tweet.
- **Region**: The geographical origin of the account that published the tweet.
- **Emotion**: The emotional category attributed to each instance, automatically determined by a pre-trained language model [70], fine-tuned with a Twitter Sentiment Analysis dataset [71].

Along with class distributions, Table 6.2 (a combination of Tables 4.9, 4.10, and 4.11) shows a significant presence of tweets originating from *America*. There is also a remarkable prevalence of messages that elicit the emotion of *fear*, which can be largely attributed to the main topics discussed on the dataset, including but not limited to the COVID-19 pandemic, the dynamics of migration, and conflicts of war.

6.2.2 Baseline

A BERT-based classifier, exclusively trained by processing the tweet in the main input without incorporating any additional information in the auxiliary input, will serve as our baseline, hereinafter referred to as **BERT-BL**.
We used BERTweet, a large-scale language model that has been pre-trained on 850 million English tweets [56], for both **BERT-BL** and **BERT-CA**. The model parameters, selected after numerous tuning iterations, consist of a batch size of 32, a learning rate of $2e-5$, an *Adam* optimizer, a max sequence length of 250 and a total of 3 epochs.

Table 6.2: Statistics of main partitions from *PropitterX* corpus

Class	**Train**	**Dev**	**Test**
Propaganda	77,167	6,454	7,045
Non-propaganda	216,313	32,057	31,456
Contextual Features			
Region			
America	201,210	22,755	22,433
Asia	26,380	5,991	6,544
Europe	18,947	4,084	3,908
Middle East	26,970	4,201	4,225
Others	19,973	1,480	1,391
Bias			
Left-wing	142,686	19,270	16,999
Right-wing	119,823	13,619	14,805
Misc.	30,971	5,622	6,697
Emotion			
Anger	47,436	5,792	5,305
Disgust	2,279	280	217
Fear	148,123	20,436	21,885
Joy	20,727	2,430	2,162
Neutral	61,400	7,766	7,194
Sadness	10,127	1,446	1,336
Surprise	3,388	361	402

6.3 Experiments

6.3.1 On the impact of adding context during the classification process.

To compare the performance of classifiers with and without contextual features and evaluate our method, several experiments were conducted using BERTweet, initially without incorporating any form of contextual attribute, i.e., relying solely on the tweet's text. Subsequently, we introduced contextual attributes

Table 6.3: F1 classification results of **BERT-CA** over the propaganda class adding different contextual features. The third column shows the probability, according to the Bayesian Wilcoxon signed-rank test, of the classifier being better than BERT-BL.

Contextual Feature Added	**F1-Prop $\pm$ Std. Dev.**	**Probability**
None (*BERT-BL*)	0.8171 ± 0.0043	-
Bias	0.8232 ± 0.0090	0.203
Region	0.8532 ± 0.0072	0.999
Emotion	0.8213 ± 0.0083	0.068
Bias + Region	$\mathbf{0.8750 \pm 0.0073}$	0.999
Bias + Emotion	0.8343 ± 0.0063	0.948
Region + Emotion	0.8473 ± 0.0054	0.996
Bias + Region + Emotion	0.8630 ± 0.0126	0.999

in BERTweet's auxiliary input, testing each feature individually, followed by combinations of two, and ultimately all three attributes together. For every classification variant, we opted to execute five runs and then calculated the average. These findings are presented in Table 6.3. As shown, all variants that incorporate context demonstrate an enhancement compared to the baseline, with the *Region* attribute performing the best when used alone. Nonetheless, the best classification results are obtained through the use of attributes in combination, with *Bias + Region* exhibiting the highest F1-score over the propaganda class, with a relative improvement of 7.08% over **BERT-BL**. For the sake of clarity, when we discuss the **BERT-CA** model in the following analyses, we specifically refer to the variant that includes all three types of context (even if it was not the configuration yielding the highest scores).

Statistical significance test

To assess the significance of incorporating context within the classification framework, we implemented a Bayesian Wilcoxon signed-rank test. This particular test constitutes a non-parametric Bayesian adaptation of the Wilcoxon signed-rank test, structured on the Dirichlet process, and it is recommended for the direct

comparison of classifiers [72], [73]. Based on the collected data, the test calculates the posterior probability for both the null and alternative hypotheses, giving a clear probability of one method outperforming the other (when evaluating two treatments), which avoids the abstract interpretations often associated with frequentist tests. As evidenced by the third column of Table 6.3, according to the Bayesian Wilcoxon signed-rank test, the likelihood that the incorporation of contextual features into the classification process yields superior scores compared to only using the texts of the tweets is greater than 94% in 5 of 7 cases (the cases where the improvements are not statistically significant correspond to adding only *bias* and only *emotion*).

6.3.1.1 Fixed and new classification mistakes by BERT-CA

Attempting to shed light on the impact and influence of incorporating contextual information for detecting propaganda, we carried out an analysis to quantify the number of errors made by **BERT-BL** that were corrected (i.e., classified correctly) by **BERT-CA**. Besides, the opposite kind of error was also analyzed. For this purpose, a total of five iterations of the baseline **BERT-BL** classifier and five iterations of the contextual-aware **BERT-CA** model were executed. Only those instances in which all five iterations of a classifier yielded identical outcomes were considered. In other terms, we omitted those occurrences in which, during the five iterations, a classifier made ambiguous predictions (for example, predicting *"propaganda"* in the first iteration and *"non-propaganda"* in the fifth iteration). Subsequently, we compared the predictions generated by both models concerning the ground truth. Following this scheme, the context-aware model successfully rectified wrong predictions made by the baseline model in a total of 205 instances (0.5% of the Test set). From them, 103 tweets were classified as *non-propaganda* and 102 as *propaganda*, a nearly perfect balance between both categories. Regarding emotional content among the 205 cases, *fear* and *anger* are

the predominant emotions observed. This finding aligns closely with the overall emotional distribution of the entire collection. With respect to *political bias*, 58 corrections pertain to *left-leaning* tweets while 105 correspond to *right-leaning* tweets. Conversely, the *contextual-aware* model introduced 28 new mistakes, i.e., it changed the label that the baseline model had accurately predicted. Among them, the ground truth label of 17 tweets was *"non-propaganda"* and 11 *"propaganda"*, with all *propaganda* tweets being of *left-wing* bias. One potential reason for this might be that the *contextual-aware* classifier observed that *left-leaning non-propaganda* was more prevalent than *left-leaning propaganda* in the train set (with a ratio of 12 to 1).

In Table 6.4, we offer a few examples of the corrections and new mistakes made by the *contextual-aware* model. In row 1, there is an instance of potential *propaganda*, where the tweet uses emotional language and relies on authority (a military veteran) to sway opinion. A correct prediction was achieved by incorporating the detail that the tweet comes from the *Middle East* region, the political orientation of the source is *right*, and the message elicits feelings of *sadness* (likely alluding to the veteran's condition). Row 2 also contains an example of potential *propaganda*. The phrasing, with words such as "unacceptable" and "reportedly", and the implicit comparison are suggestive of a biased viewpoint. When context about the tweet is added to BERT-CA, the prediction is fixed. Row 3 shows an instance where the core message is factual but could be framed to emphasize disruption or China's COVID policies. It depends on context and whether the URL leads to a biased source. By noting that it originates from a source that is neither *left-leaning* nor *right-leaning*, that it is based in *America*, and that the statement incited *fear* (likely referring to the mention of testing positive for COVID), the contextual model's assessment shifted correctly to *non-propaganda*. In row 4, the message appears to be sensationalistic, but not inherently propagandistic, as the baseline model predicted. A correct prediction was achieved by taking into

account that the tweet is *left-leaning*, comes from *America* and evokes *joy*. In instances where the contextual model makes mistakes, row 5 shows a message that uses emotionally charged language ("gem," "precious") and a loaded term ("redress") to persuade without full explanation, characteristic of propaganda. By adding context, the model found no propaganda in the message. Row 6 illustrates an example where the phrasing is neutral, but the underlying fear mentioned could be manipulated or exaggerated. The predominant emotion identified as *fear*, may have caused the model to mistakenly interpret it as *propaganda*.

6.3.2 On the impact of adding context when using limited training data

Data for propaganda detection can be scarce. Accordingly, we aimed to investigate the usefulness of contextual attributes in scenarios where there are only a few labeled instances. Our intuition is that, even with restricted training resources, contextual features can effectively aid in differentiating between propagandistic and non-propagandistic content. For that purpose, we replicated our first experiment with **BERT-CA** incorporating all three types of context available, altering the volume of data used for training the classifiers, effectively halving the training set up to 4 times. The findings from this experiment are illustrated in Figure 6.2, where it is evident that the value of integrating context into the classification process remains to some degree unchanged in relation to the baseline model as the amount of training data diminishes. In fact, the performance of the contextual model decreases slightly less than that of the baseline model when the training data is at its lowest. For example, it is worth noting that the gap in F1-score between the two models widens, going from a relative difference of 4.8% achieved with the complete training set to 7.8% when using the smaller training data volume. This indicates that, in circumstances characterized by a limited

Table 6.4: Examples of fixed and new mistakes by adding contextual features to the classifier. A label of 1 means *propaganda*, while 0 means *non-propaganda*.

Tweet	**Label**	**BERT-BL**	**Context Added**	**BERT-CA**	**Mistake**
A British veteran who lost both his legs in an explosion while serving in Afghanistan describes the situation in the country as "shameful" URL	1	0	Right, Middle-East, Sadness	1	Fixed
Cotton, an Afghan war vet, said it is unacceptable that US forces are not helping Americans get to the Kabul airport when the British and French forces reportedly are aiding their citizens. URL	1	0	Right, America, Fear	1	Fixed
China's Ningbo-Zhoushan container port, the world's third-busiest, remained partially closed for a sixth day following its halt of all inbound and outbound container services at its Meishan terminal after one employee tested positive for the coronavirus. URL	0	1	Neutral (Least-Biased), America, Fear	0	Fixed
Experienced' Sloth Mom Lunesta Gives Birth to Her Fifth Baby at New England Zoo URL	0	1	Left, America, Joy	0	Fixed
Lifta must be saved not only because it is a gem of precious natural beauty and human architecture, but also because it is a step towards redress. URL	1	1	Left, America, Anger	0	New
The announcement follows fears that Ukraine would ban the pilgrimage for a second year due to the COVID-19 pandemic URL	0	0	Unknown, Others, Fear	1	New

Figure 6.2: Average classification scores (F1 over the propaganda class), incorporating contextual features and changing the volume of train data.

availability of labeled training data, the incorporation of contextual attributes can play a significant role in "constraining" particular instances of propaganda. By providing insights, such as identifying the source from which it emanates, context can consequently facilitate a better differentiation between the classes of *propaganda* and *non-propaganda*.

6.3.3 Classifying Tweets from Unknown Sources

Accounts linked to various news media organizations, government entities, and political parties—whose contextual information is often well-documented—are primarily responsible for the creation and dissemination of propaganda. However, the accessibility and versatility of social media platforms allow for a diverse type of

casual users to also engage in the creation and sharing of propagandist messages. This situation can lead to the context surrounding these users being unclear or entirely unknown. We encountered the challenge of addressing situations in which such contextual attributes are absent, specifically how to navigate the scenario where a prediction about a tweet must be made without knowledge of its source of origin. Considering that prior experiments used a pre-trained language model to identify *emotions* evoked by texts, we aimed to explore the feasibility of training models to recognize both the *political bias* within the message as well as its *geographical origin*. The current experiment involves the use of the *Train*, *Dev*, and *Test* partitions of the *PropitterX* dataset. However, rather than employing *propaganda* or *non-propaganda* as the target labels for training a BERTweet model, we designated the *bias* attribute as the target. Additionally, in a separate experiment, we applied the same concept but aimed to predict the *region* as the target instead. According to Table 6.5, predicting the *region* poses a greater challenge than *bias*, primarily due to the fact that this attribute is categorized into a larger number of classes (a total of 5), four of which are minority classes when compared to the volume of tweets originating from *America*. In terms of *bias*, the models were able to achieve F1-scores of 0.75 and 0.77 for *left* and *right biases*, respectively. Making predictions over the *miscellaneous biases*, a minority class, proved to be more challenging.

Leveraging these predicted attributes (hereafter referred to as *calculated contextual features*), we executed five classification runs using all three types of attributes (*bias*, *region*, and *emotion*), and other additional five runs employing solely the calculated *bias* (as it was easier to predict accurately in contrast to *region*). Consequently, we noted that incorporating all calculated features resulted in a decline in classification performance when compared to the baseline model. Conversely, using the most effective available calculated attribute in isolation as the only context added marginally surpassed the baseline, albeit without achiev-

Table 6.5: Classification results obtained by predicting the bias and the region of the tweets in the test set. "Support" indicates the number of instances corresponding to each class.

	Precision	Recall	F1-score	Support
Left Bias	0.6418	0.9203	0.7562	16999
Right Bias	0.8763	0.6931	0.7740	14805
Misc. Biases	0.6663	0.2403	0.3532	6697
America	0.8536	0.9479	0.8983	22433
Asia	0.9468	0.4080	0.5703	6544
Middle East	0.5187	0.8293	0.6383	4225
Europe	0.7352	0.5448	0.6258	3908
Others	0.5769	0.4637	0.5141	1391

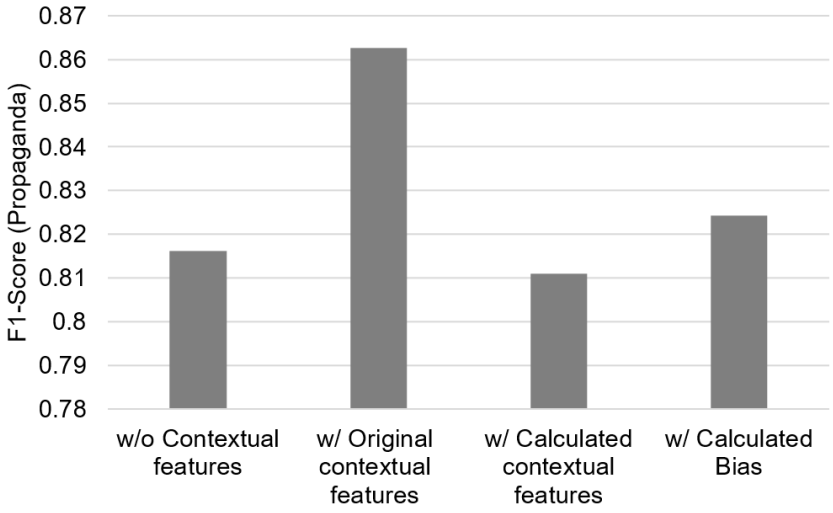

Figure 6.3: Average classification scores (F1 over the propaganda class) obtained by incorporating contextual features as a secondary input.

ing statistical significance. A comparison between using the original contextual features vs calculated ones is shown in Figure 6.3.

6.3.4 Classifying Tweets from Diplomatic Profiles and Government Authorities

The studies conducted with *PropitterX* were aimed at detecting propaganda in tweets from news organizations. In order to further evaluate the effectiveness of our classification strategy, we carried out an experiment considering a different type of propaganda, spread through accounts associated with government entities.

In 2023, *DIPROMATS* [12] introduced datasets containing propaganda from Twitter accounts of diplomats, ambassadors, and governmental entities (along with information about the account's *country of origin* and *tweet type*). They are collections of tweets published by Chinese, Russian, European Union, and United States authorities between January 2020 and March 2021. Table 6.6 shows the distribution of the datasets for both English and Spanish (refer to [12] for more information). Upon conducting a comparative analysis between *PropitterX* and *DIPROMATS*, it becomes evident that the *country* attribute is more evenly distributed in *DIPROMATS*, and while the *political bias* attribute is absent, the *type of tweet* serves as a new contextual feature (indicating whether the post was a tweet, retweet, reply, or quote). Regarding emotions (inferred via [70], [74]), we observe that in *DIPROMATS*, there is a tendency for messages issued by diplomats in English to lean towards a more positive tone, with *joy* being the most frequently evoked emotion.

To evaluate our approach, we concentrated on the task of *binary propaganda identification* in both English and Spanish, using BERTweet and RoBERTuito [75], respectively. In both models, the same hyperparameter values described in Section 6.2 were applied. Table 6.7 presents the results of our methodology. Our approach, which incorporates context in the same manner we have delineated in Section 6.1, yielded the best results in Spanish by integrating the *type of tweet* as contextual information into the auxiliary input of the RoBERTuito model, and

Table 6.6: Data distribution for the English and Spanish corpora.

Class	**Train ENG**	**Test ENG**	**Train SPA**	**Test SPA**
Propaganda	1,974	N/A	1,199	N/A
Non-propaganda	6,434	N/A	4,921	N/A
	Contextual Features			
Country				
China	2,170	852	2,178	819
European Union	2,043	873	1,508	957
Russia	2,005	955	795	596
USA	2,190	924	1,639	1,099
Type of tweet				
Tweet	6,742	2,856	3,586	2,302
Quoted	825	356	888	541
Retweet	473	227	1,221	401
Reply	368	165	425	227
Emotion*				
Anger	2,270	760	259	90
Fear	276	72	5	4
Joy	5,216	2,569	649	376
Love	114	53	N/A	N/A
Others	N/A	N/A	4,961	2,919
Sadness	508	141	224	66
Surprise	24	9	22	16

*Inferred, not available in the original dataset.

in English by incorporating the *emotion* evoked by the message as a contextual token. Clearly, the incorporation of contextual attributes enhances the performance of our classifiers, with F1-score gains of up to 3.5 points when evaluated with the Spanish dataset, and 0.96 points in the English dataset. It is significant to emphasize that the baseline models, devoid of context, were predominantly ranked lower than our contextualized models.

Our best results averaging both languages scored a 0.7953 F1-macro. For comparison, we offer a summary of other approaches that have been tested in the same dataset. A research team developed a hierarchical model to detect and

characterize propaganda techniques in text [76]. Their methodology involved fine-tuning a XLM-RoBERTa model using multiple datasets, achieving a F1-score of 0.7770. The strategy employed in [77] assessed linguistic attributes and sentence embeddings derived from various LLMs, encompassing models tailored for English, Spanish, and additional multilingual frameworks, resulting in a F1-score of 0.7734. A different research team achieved a F1-score of 0.7732, implementing data augmentation techniques to enhance the sample size by translating Spanish samples into English and the other way around [78]. They used BETO for tweets written in Spanish and a RoBERTa-based variant of TimeLM [79] for tweets in English, fine-tuning with a Discrepancy Correction Procedure to avoid inconsistencies in labeling.

In terms of propaganda detection results only in DIPROMATS-English, our best *context-aware* model scored a 0.8062 F1-macro. For comparison, the strategy submitted by [65], "injecting" geographical context to each message, achieved a 0.8011 F1-macro. The approach by [64] which "augmented" the original messages with emotions, sentiments, interactions, and countries, achieved a 0.7953 F1-macro.

The boxplots in Figure 6.4 provide a visual representation of the comparison between our context-aware approach and other classification strategies tested by other studies for DIPROMATS 2023, measured in terms of F1-macro scores. Our approach is prominently positioned at the very limits of the upper whiskers, which means that our results are indeed part of the highest scores achieved using that dataset.

In terms of statistical significance, we implemented the Bayesian Wilcoxon signed-rank test within a 5-fold cross-validation framework using the training set across both languages. Despite the observation of improved classification scores when contextual information is incorporated, the statistical analysis concluded

Table 6.7: Official results obtained by our submissions in the DIPROMATS 2023 shared task, corresponding to **BERT-BL** and **BERT-CA**.

Task 1	Contextual Feature Added	F1-score	Rank
SPA	None	0.7730	10 of 18
	Type of Tweet	**0.8089**	**1 of 18**
ENG	None	0.7966	6 of 30
	Emotion	**0.8062**	**2 of 30**
AVG	None	0.7880	4 of 16
	Emotion	**0.7953**	**1 of 16**

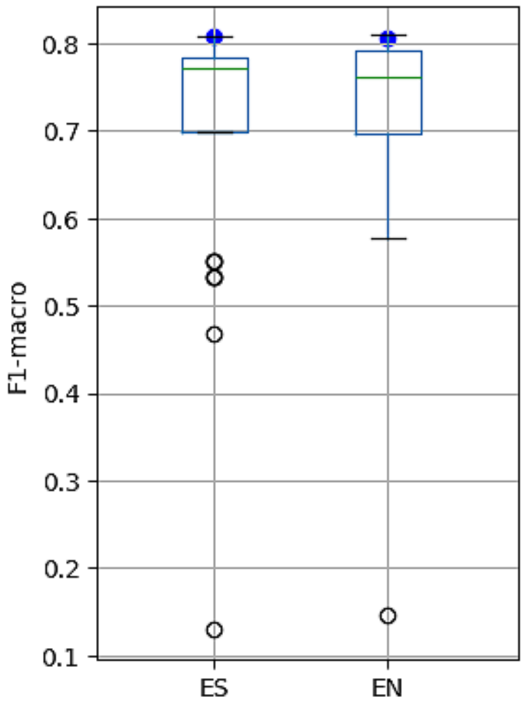

Figure 6.4: Box plots of the results for Task 1 of *DIPROMATS* 2023. The blue dots represent our best scores using contextual features in the shared task.

that the advantage of the *contextual-aware* model was not significant for English, but significant in the Spanish language (see Table 6.8).

We speculate that the use of data in English and the source of the tweets (coming from official government sources and entities) generated a complex scenario, in which the contextual attributes might not have provided sufficient infor-

Table 6.8: Bayesian Signed-Rank Test results applied in *DIPROMATS* Spanish Train set. **BERT-BL** only takes into account tweets' text, **BERT-CA** uses both text and contextual features.

	Probability
BERT-BL >BERT-CA	0.199
Region of Practical Equivalence	0.199
BERT-CA >BERT-BL	0.602

mation to achieve a better distinction between *propaganda* and *non-propaganda* classes. We do not attribute this result exclusively to the source of the data (since statistical significance was achieved in Spanish), but rather the fact that there may be contextual differences between languages.

6.4 Summary

We assessed the usefulness of incorporating contextual information for performing propaganda detection on tweets. We took advantage of corpora in the state-of-the-art having posts annotated on the presence of propaganda which also have attributes regarding the context surrounding the messages like political affiliation of the source, the geographical location from which the message was posted, the emotion evoked by them, and the type of tweets. These features allowed us to conduct experiments under diverse scenarios: without using any context, adding different combinations of contextual aspects, and assessing the usefulness of context on data scarcity. Furthermore, we evaluated the possibility of forecasting the contextual attributes considering this information is unavailable. We observed that enhancing the textual content of the posts through contextual information improves the classification rates, with F1-score gains of up to 5.8 points in *PropitterX*, 3.1 points in *DIPROMATS-Spanish* and 0.9 points in *DIPROMATS-English*. On the other hand, our experiments using the *context-aware* approach

under scenarios in which contextual information is initially unavailable showed only marginal improvements in the detection scores; these improvements did not reach a level of statistical significance that would validate our findings in a broader context.

Chapter 7

Conclusions and Future Work

Revisiting our Research Questions

In this section, we will address the research questions outlined at the beginning of this document.

How can resources derived from news articles be leveraged to identify computational propaganda on Twitter?
In the development of the *Propitter* dataset, we integrated two methodologies for dataset construction observed in previous propaganda-related studies: pseudo-labeling through distant supervision, and manually annotated data used to refine and filter the data we collected. This process involved using an existing data collection comprising sentences from news articles that had been manually labeled[28]. With this dataset, we trained multiple classifiers to make class predictions (*propaganda* or *non-propaganda*) and to evaluate whether these predictions aligned with the pseudo-labels generated through distant supervision. This approach enabled us to enhance the quality of our dataset beyond what could have been achieved by relying exclusively on distant supervision. Overall, the endeavor

of detecting computational propaganda is enhanced by the development of new resources that encompass diverse sources and domains beyond mere news articles. To achieve this, we have leveraged existing resources and integrated various labeling methods to introduce a novel corpus (*PropitterX*).

What are the differences (in terms of topics covered, emotions evoked) in computational propaganda from tweets based on its context?
As we remember from the conclusion of Chapter 5, some intriguing observations regarding the link between propaganda and various contextual elements indicate that:

- Propaganda generated by sources with a left-leaning bias is distinct from that created by those with a right-leaning bias, highlighting its significant impact on the message content. For instance, we noted that leftist propaganda predominantly concentrated on the pandemic and global issues when addressing specific regions of the world, like the US and China. In contrast, right-wing propaganda was centered on matters related to Joe Biden, abortion, and racial issues.

- Trending topics linked to propaganda appear to be dynamic, affecting the efficacy of the systems designed to recognize such messages. For instance, we noted a shift in the frequency of the usage of terms like *"Trump"*, *"COVID"*, and *"vaccine"* during a specific period of time. However, an event involving different terms, such as *"military"*, *"Taliban"*, and *"Afghanistan"*, can rise in prominence and momentarily overshadow other topics, fundamentally altering what is deemed relevant for a classifier tasked with detecting propaganda.

- Neutral propaganda content encompasses a wider array of subjects compared to propaganda centered around specific emotions. The two most prominent emotions in PropitterX are *fear* and *anger*. Fear-driven pro-

paganda explores issues related to warfare and public health (including pandemics, vaccines, and COVID). Anger-inducing propaganda addresses topics linked to racism and immigration. While neutral tweets containing propaganda use some of the same terms as these emotions, they offer a more consistent narrative where no specific subjects stand out.

- There exists a diversity of propaganda across various geographical areas. Each area within our dataset concentrated on distinct subjects. Notably, it is observed how each area addresses matters concerning politics and government, although the names of the places and leaders mentioned in the tweets vary significantly. The only regions where specific common propaganda terms are recognized are *"America"* and *"Others"* (with a notable frequency of mentions of D. Trump and J. Biden).

How can contextual information of messages be incorporated to improve the effectiveness of propaganda detection in them?

Our proposed *context-aware* approach is based on BERT models. The input representation of the model is designed to accommodate not only a single text sentence but also a combination of two text sequences. We chose to leverage this to provide context to the sentence presented to the main text input for the task of detecting propaganda. In this manner, the incorporation of contextual attributes within propaganda classifiers resulted in an increase in detection performance, with some instances showing statistically significant improvements over baseline results. Notably, the addition of contextual attributes to texts was accomplished without altering the architecture of BERT-based classifiers, which are renowned for their state-of-the-art performance. This renders our method both straightforward to implement and highly competitive. Analysis of data from collections *PropitterX* and *DIPROMATS* demonstrated that even the use of a single contextual attribute (as opposed to combinations) yields superior results compared to classification performed without this information.

Final Remarks

As a starting point for this research endeavor, we identified a substantial gap in the exploration and analysis concerning the phenomenon of propaganda as disseminated through social network platforms by various news media organizations, which was largely overlooked in previous studies. To address this, the primary contribution of this work was the creation of a corpus, specifically designed for the purpose of facilitating a deeper understanding of this complex issue.

Our hypothesis (based on one of our research questions) was that contextual information of messages can be incorporated in the classification process to improve the effectiveness of propaganda detection.

In this research, we have investigated the identification of computational propaganda within a social network by leveraging various contextual attributes. Each contextual attribute can be linked to one or multiple propaganda techniques:

- Political bias significantly impacts and reveals itself through *"Name calling or labeling"* and *"Slogans"*.
- Geographical origin is tied to *"Flag-waving"*.
- Emotions serve as a crucial element in techniques such as *"Loaded Language"* and *"Appeal to fear/prejudice"*.
- One potential connection between propaganda techniques and tweet types lies in the importance of recognizing whether a message is aimed directly at another account (to provoke *"Name calling"* or *"Doubt"*), merely retweeted/repeated (*"Repetition"*), or referencing other sources (*"Appeal to authority"*). These factors can be pivotal when attempting to steer the direction of a discourse or argument.

Our proposal for a *Contextual-Aware Approach* incorporates these contextual attributes into the training and classification processes of propaganda. The findings from our experiments indicate that considering all these various types of contextual attributes outperforms baseline strategies that ignore this additional information, with our method being exceptionally competitive and obtaining the highest scores when compared to other classification methods showcased at DIPROMATS 2023 workshop. Our analysis of the *contextual-approach* reveals that integrating this contextual information is crucial for detecting propaganda, as it is vital to comprehend its origins, the political bias it is linked to, and the emotion it aims to provoke in the reader.

Scope and Limitations

- This research concentrated on enhancing the detection of computational propaganda disseminated on Twitter. Consequently, it is important to understand that the conclusions we have drawn from our research may not be universally applicable or relevant to other social networks, given the unique characteristics and dynamics that each platform possesses.

- Since the tweets that have been analyzed for our research purposes are mostly written in the English language, this consequently means that our outcomes and the findings derived from them are inherently restricted to this particular language.

- Due to the way we partitioned Propitter, news sources that are in the training partition are also in the development and test partitions. This presents a potential case of data leakage, which we mitigated by taking tweets from a sizable pool of news sources (244).

- Our study does encounter certain limitations, particularly regarding the

labeling of the *PropitterX* dataset. As with any endeavor involving the creation of a data collection, the task of labeling data is complex. We acknowledge that our corpus is subject to the same challenges. This limitation must be carefully considered, particularly given that the labeling process inherently involves several assumptions. In the context of our research, we aimed to integrate the strengths of labeling approaches observed in previous propaganda-related works. Our labeling process merges the advantages of distant supervision with a filter that has been trained using externally sourced instances that were manually labeled in related studies. Consequently, we strongly encourage anyone intending to use our data or reference our findings to carefully consider the conditions under which our results were derived.

- Although it is true that the instances within our dataset were not individually reviewed to verify or disprove the presence of propaganda, we had the opportunity to test our classification approach across multiple datasets (namely *DIPROMATS English* and *Spanish*), where, to the best of our knowledge, manual labeling was conducted. Having tested our model on the *DIPROMATS* collection, and as we described before in Section 6.3.4, our approach turned out to be competitive against other classification strategies, some of them even also making use of contextual attributes in a different way than we proposed.

Social Concerns

The significance of data management and artificial intelligence within the field of social studies is well recognized. Our research exclusively used data derived from news media sources, the publications of which were publicly available during the period of data acquisition. Nevertheless, it is important to emphasize that we

assert all data, models and conclusions derived from this work are intended solely for research purposes, and not for any unethical uses.

Future Work

- As future work, we aim to investigate additional types of contextual attributes that could enhance the classification process. Among our suggestions is the inclusion of whether messages feature multimedia elements at the time they are shared on social networks, as well as the degree of engagement those messages receive.

- We intend to test our methodology with more data collections as they become available in the future. This will also entail applying our approach to data from different domains, including social networks other than Twitter.

- If additional resources become available in the future, we would be eager to evaluate our strategy in languages beyond English and Spanish, and even tailor the methodology for different classification tasks, incorporating pertinent information for each one of them as types of "context".

- We would like to explore the application of *Large Language Models* (LLMs) in the detection of propaganda. Given the rapid advancements in NLP and the increasing capabilities of LLMs to understand and generate human language, these models present a promising tool for analyzing and identifying propagandistic content. The understanding of language, context, and sentiment demonstrated by LLMs could be leveraged to detect patterns and manipulative techniques commonly used in propaganda. This would not only contribute to the growing body of research on automated content analysis but also offer practical insights into the potential for LLMs to

serve as instruments in combating misinformation in an increasingly digital world.

Bibliography

[1] G. Meikle, *Social Media: Communication, Sharing and Visibility*. Routledge, 2016, ISBN: 9780415712231. [Online]. Available: https://books.google.com.mx/books?id=H-XXsgEACAAJ.

[2] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Systems with Applications*, vol. 153, p. 112 986, 2020, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2019.112986. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417419307043.

[3] C. Miller, *How to Detect and Analyze Propaganda ...: An Address Delivered at Town Hall, Monday, February 20, 1939* (A Town Hall pamphlet). Town Hall, Incorporated, 1939. [Online]. Available: https://books.google.com.mx/books?id=UAc4AAAAMAAJ.

[4] N. Newman, W. Dutton, and G. Blank, "Social media in the changing ecology of news: The fourth and fifth estate in britain," *International Journal of Internet Science*, vol. 7, Jul. 2012.

[5] J. Tucker, A. Guess, P. Barberá, *et al.*, "Social media, political polarization, and political disinformation: A review of the scientific literature," *SSRN Electronic Journal*, Jan. 2018. DOI: 10.2139/ssrn.3144139.

[6] J. A. Tucker, Y. Theocharis, M. E. Roberts, and P. Barberá, "From liberation to turmoil: Social media and democracy," *Journal of democracy*, vol. 28, no. 4, pp. 46–59, 2017.

[7] G. Da San Martino, S. Cresci, A. Barrón-Cedeño, S. Yu, R. D. Pietro, and P. Nakov, "A survey on computational propaganda detection," in *IJCAI*, 2020.

[8] R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 6086–6093.

[9] A. Sardo, "Categories, balancing, and fake news: The jurisprudence of the european court of human rights," *Canadian Journal of Law & Jurisprudence*, vol. 33, no. 2, pp. 435–460, 2020. DOI: 10.1017/cjlj.2020.5.

[10] L. Wang, X. Shen, G. de Melo, and G. Weikum, "Cross-domain learning for classifying propaganda in online contents," in *Truth and Trust Online Conference*, Hacks Hackers, 2020, pp. 21–31.

[11] G. Bolsover and P. Howard, "Computational propaganda and political big data: Moving toward a more critical research agenda," *Big data*, vol. 5, no. 4, pp. 273–276, 2017.

[12] Pablo Moral, Guillermo Marco, Julio Gonzalo, Jorge Carrillo-de-Albornoz, and Iván Gonzalo-Verdugo, "Overview of DIPROMATS 2023: Automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers," *Procesamiento del Lenguaje Natural*, vol. 71, Sep. 2023.

[13] S. Marsland, *Machine Learning: An Algorithmic Perspective*, 2nd Ed. Chapman and Hall/CRC, 2014, 457 Pages, ISBN: 978-1466583283.

[14] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow (2019, O'reilly)*. O'Reilly Media, 2017, ISBN: 9781492032649.

[15] D. Jurafsky and J. H. Martin, *Speech and Language Processing (2nd Edition)*. USA: Prentice-Hall, Inc., 2009, ISBN: 0131873210.

[16] Q. Li, H. Peng, J. Li, *et al.*, "A survey on text classification: From traditional to deep learning," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, Apr. 2022, ISSN: 2157-6904. DOI: 10.1145/3495162. [Online]. Available: https://doi.org/10.1145/3495162.

[17] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008, ISBN: 0521865719. [Online]. Available: https://nlp.stanford.edu/IR-book/.

[18] C. Aggarwal, *Machine Learning for Text*. Springer International Publishing, 2018, ISBN: 9783319735313. [Online]. Available: https://books.google.com.mx/books?id=uVNSDwAAQBAJ.

[19] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.

[20] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China technological sciences*, vol. 63, no. 10, pp. 1872–1897, 2020.

[21] Z. Wan, "Text Classification: A Perspective of Deep Learning Methods," *arXiv preprint arXiv:2309.13761*, 2023.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://www.aclweb.org/anthology/N19-1423.

[23] R. Dror, L. Peled-Cohen, S. Shlomov, and R. Reichart, *Statistical Significance Testing for Natural Language Processing*. Morgan & Claypool, 2020, ISBN: 1681737957.

[24] C. Wardle, "FIRST DRAFT'S Essential Guide to Understanding information disorder," *First Draft*, no. October, p. 61, 2019. [Online]. Available: https://firstdraftnews.org/wp-content/uploads/2019/10/Information%7B%5C_%7DDisorder%7B%5C_%7DDigital%7B%5C_%7DAW.pdf?x47711.

[25] M. Ginsberg, "8 - propaganda art as a powerful weapon for promoting nationalism, patriotism and hatred towards the enemy," in *Inside the World's Major East Asian Collections*, ser. Chandos Information Professional Series, A. Cho, P. Lo, and D. K. Chiu, Eds., Chandos Publishing, 2017, pp. 85–95, ISBN: 978-0-08-102145-3. DOI: https://doi.org/10.1016/B978-0-08-102145-3.00008-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978008102145300008X.

[26] B. Sinno, B. Oviedo, K. Atwell, M. Alikhani, and J. J. Li, "Political Ideology and Polarization: A Multi-dimensional Approach," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 231–243. DOI: 10.18653/v1/2022.naacl-main.17. [Online]. Available: https://aclanthology.org/2022.naacl-main.17/.

[27] J. Sproule, D. Culbert, G. Jowett, and K. Short, *Propaganda and Democracy: The American Experience of Media and Mass Persuasion* (Cambridge Studies in the History of Mass Communication). Cambridge University Press, 1997, ISBN: 9780521470223. [Online]. Available: https://books.google.com.mx/books?id=Xv9cXHL9f18C.

[28] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov, "Fine-grained analysis of propaganda in news article," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5636–5646. DOI: 10.18653/v1/D19-1565. [Online]. Available: https://www.aclweb.org/anthology/D19-1565.

[29] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2931–2937. DOI: 10.18653/v1/D17-1317. [Online]. Available: https://www.aclweb.org/anthology/D17-1317.

[30] D. Graff and C. Cieri, *English Gigaword LDC2003T05*, https://catalog.ldc.upenn.edu/LDC2003T05, Web Download. Philadelphia: Linguistic Data Consortium, Jan. 2003.

[31] A. Barrón-Cedeño, I. Jaradat, G. Da San Martino, and P. Nakov, "Proppy: Organizing the news based on their propagandistic content," *Information Processing & Management*, vol. 56, May 2019. DOI: 10.1016/j.ipm.2019.03.005.

[32] B. Horne, S. Khedr, and S. Adali, "Sampling the news producers: A large news and feature data set for the study of the complex media landscape," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, 2018.

[33] G. Da San Martino, A. Barrón-Cedeño, and P. Nakov, "Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection," in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 162–170. DOI: 10.18653/v1/D19-

5024. [Online]. Available: https://www.aclweb.org/anthology/D19-5024.

[34] A. Fadel, I. Tuffaha, and M. Al-Ayyoub, "Pretrained ensemble learning for fine-grained propaganda detection," in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 139–142. DOI: 10.18653/v1/D19-5020. [Online]. Available: https://aclanthology.org/D19-5020.

[35] D. Cer, Y. Yang, S.-y. Kong, *et al.*, "Universal sentence encoder for English," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 169–174. DOI: 10.18653/v1/D18-2029. [Online]. Available: https://aclanthology.org/D18-2029.

[36] G. Da San Martino, A. Barrón-Cedeno, H. Wachsmuth, R. Petrov, and P. Nakov, "Semeval-2020 task 11: Detection of propaganda techniques in news articles," in *Proceedings of the fourteenth workshop on semantic evaluation*, 2020, pp. 1377–1414.

[37] G. Morio, T. Morishita, H. Ozaki, and T. Miyoshi, "Hitachi at SemEval-2020 task 11: An empirical study of pre-trained transformer family for propaganda detection," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds., Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1739–1748. DOI: 10.18653/v1/2020.semeval-1.228. [Online]. Available: https://aclanthology.org/2020.semeval-1.228/.

[38] B. Polonijo, S. Šuman, and I. Šimac, "Propaganda detection using sentiment aware ensemble deep learning," in *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, 2021, pp. 199–204. DOI: 10.23919/MIPRO52101.2021.9596654.

[39] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[40] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, 2014, pp. 216–225.

[41] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM '11, Hong Kong, China: Association for Computing Machinery, 2011, pp. 177–186, ISBN: 9781450304931. DOI: 10.1145/1935826.1935863. [Online]. Available: https://doi.org/10.1145/1935826.1935863.

[42] J. W. Pennebaker, M. E. Francis, and R. J. Booth, *Linguistic Inquiry and Word Count: LIWC2001*. Lawrence Erlbaum Associates, 2001.

[43] P. Vijayaraghavan and S. Vosoughi, "TWEETSPIN: Fine-grained propaganda detection in social media using multi-view representations," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 3433–3448. DOI: 10.18653/v1/2022.naacl-main.251. [Online]. Available: https://aclanthology.org/2022.naacl-main.251.

[44] S. M. Jiménez-Zafra, F. Rangel, and M. M.-y. Gómez, "Overview of iberlef 2023: Natural language processing challenges for spanish and other iberian languages," in *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEURWS. org*, 2023.

[45] P. Moral, G. Marco, J. Gonzalo, J. Carrillo-de-Albornoz, and I. Gonzalo-Verdugo, "Overview of DIPROMATS 2023: Automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers," *Procesamiento del lenguaje natural*, vol. 71, pp. 397–407, 2023.

[46] O. Balalau and R. Horincar, "From the stage to the audience: Propaganda on reddit," *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 3540–3550, 2021. DOI: 10.18653/v1/2021.eacl-main.309.

[47] D. Dimitrov, B. B. Ali, S. Shaar, *et al.*, "Semeval-2021 task 6: Detection of persuasion techniques in texts and images," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 2021, pp. 70–98.

[48] I. Vogel and M. Meghana, "Detecting fake news spreaders on twitter from a multilingual perspective," *Proceedings - 2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020*, pp. 599–606, 2020. DOI: 10.1109/DSAA49011.2020.00084.

[49] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Suntec, Singapore: Association for Computational Linguistics, Aug. 2009, pp. 1003–1011. [Online]. Available: https://aclanthology.org/P09-1113.

[50] Y. Hua, "Understanding BERT performance in propaganda analysis," pp. 135–138, 2019. DOI: 10.18653/v1/d19-5019. arXiv: 1911.04525.

[51] S. Zhang and M. Kejriwal, "Concept drift in bias and sensationalism detection: An experimental study," in *2019 IEEE/ACM International Conference on Ad-*

vances in Social Networks Analysis and Mining (ASONAM), 2019, pp. 601–604. DOI: 10.1145/3341161.3343690.

[52] L. Rocha, F. Mourão, A. Pereira, M. A. Gonçalves, and W. Meira, "Exploiting temporal contexts in text classification," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ser. CIKM '08, Napa Valley, California, USA: Association for Computing Machinery, 2008, pp. 243–252, ISBN: 9781595939913. DOI: 10.1145/1458082.1458117. [Online]. Available: https://doi.org/10.1145/1458082.1458117.

[53] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" In *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10, Raleigh, North Carolina, USA: Association for Computing Machinery, 2010, pp. 591–600, ISBN: 9781605587998. DOI: 10.1145/1772690.1772751. [Online]. Available: https://doi.org/10.1145/1772690.1772751.

[54] D. McClosky, E. Charniak, and M. Johnson, "Effective Self-Training for Parsing," in *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, R. C. Moore, J. Bilmes, J. Chu-Carroll, and M. Sanderson, Eds., New York City, USA: Association for Computational Linguistics, Jun. 2006, pp. 152–159. [Online]. Available: https://aclanthology.org/N06-1020.

[55] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Machine Learning: ECML-98*, C. Nédellec and C. Rouveirol, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 137–142, ISBN: 978-3-540-69781-7.

[56] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English Tweets," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 9–14.

[57] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, R. Mooney, C. Brew, L.-F. Chien, and K. Kirchhoff, Eds., Vancouver, British Columbia, Canada: Association for Computational Linguistics, Oct. 2005, pp. 347–354. [Online]. Available: https://aclanthology.org/H05-1044.

[58] A. F. Cruz, G. Rocha, and H. L. Cardoso, "On document representations for detection of biased news articles," in *Proceedings of the 35th annual ACM symposium on applied computing*, 2020, pp. 892–899.

[59] P. Ekman, "Universals and cultural differences in facial expressions of emotion.," in *Nebraska symposium on motivation*, University of Nebraska Press, 1971.

[60] R. Michael and B. Breaux, "The relationship between political affiliation and beliefs about sources of "fake news"," *Cognitive Research: Principles and Implications*, vol. 6, Dec. 2021. DOI: 10.1186/s41235-021-00278-1.

[61] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[62] L. Rocha, F. Mourão, A. Pereira, M. A. Gonçalves, and W. Meira, "Exploiting temporal contexts in text classification," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ser. CIKM '08, Napa Valley, California, USA: Association for Computing Machinery, 2008, pp. 243–252, ISBN: 9781595939913. DOI: 10.1145/1458082.1458117. [Online]. Available: https://doi.org/10.1145/1458082.1458117.

[63] T. Salles, L. Rocha, G. L. Pappa, F. Mourão, W. Meira, and M. Gonçalves, "Temporally-aware algorithms for document classification," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development*

in Information Retrieval, ser. SIGIR '10, Geneva, Switzerland: Association for Computing Machinery, 2010, pp. 307–314, ISBN: 9781450301534. DOI: 10.1145/1835449.1835502. [Online]. Available: https://doi.org/10.1145/1835449.1835502.

[64] A. Pritzkau, "Investigating Propaganda Considering the Discursive Context of Utterances," in *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023)*, ser. CEUR Workshop Proceedings, vol. 3496, Jaén, Spain: CEUR-WS.org, 2023. [Online]. Available: https://ceur-ws.org/Vol-3496/dipromats-paper1.pdf.

[65] F. Jáñez-Martino and A. Barrón-Cedeño, "Unileon-UniBO at IberLEF 2023 Task DIPROMATS: RoBERTa-based Models to Climb Up the Propaganda Tree in English and Spanish," in *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023*, ser. CEUR Workshop Proceedings, vol. 3496, Jaén, Spain: CEUR-WS.org, 2023. [Online]. Available: https://ceur-ws.org/Vol-3496/dipromats-paper3.pdf.

[66] M. Casavantes, M. Montes-y-Gómez, D. I. H. Farías, L. C. González-Gurrola, and A. Barrón-Cedeño, "PropaLTL at DIPROMATS: Incorporating Contextual Features with BERT's Auxiliary Input for Propaganda Detection on Tweets," in *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023*, M. Montes-y-Gómez, F. Rangel, S. M. J. Zafra, *et al.*, Eds., ser. CEUR Workshop Proceedings, vol. 3496, Jaén, Spain: CEUR-WS.org, 2023. [Online]. Available: https://ceur-ws.org/Vol-3496/dipromats-paper2.pdf.

[67] S. Yu, J. Su, and D. Luo, "Improving bert-based text classification with auxiliary sentence and domain knowledge," *IEEE Access*, vol. 7, pp. 176 600–176 612, 2019. DOI: 10.1109/ACCESS.2019.2953990.

[68] F. Sánchez-Vega and A. P. López-Monroy, "BERT's Auxiliary Sentence focused on Word's Information for Offensiveness Detection.," in *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021)*, ser. CEUR Workshop Proceedings, vol. 2943, CEUR-WS.org, 2021. [Online]. Available: https://ceur-ws.org/Vol-2943/meoffendes_paper4.pdf.

[69] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 380–385. DOI: 10.18653/v1/N19-1035. [Online]. Available: https://aclanthology.org/N19-1035.

[70] *Model Description,bert-base-uncased-emotion*, https://huggingface.co/bhadresh-savani/bert-base-uncased-emotion, [Online; accessed 30-May-2023], 2018.

[71] *"Emotion", Dataset Summary*, https://huggingface.co/datasets/philschmid/emotion, [Online; accessed 30-May-2023], 2022.

[72] A. Benavoli, F. Mangili, G. Corani, M. Zaffalon, and F. Ruggeri, "A bayesian wilcoxon signed-rank test based on the dirichlet process," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14, Beijing, China: JMLR.org, 2014, II–1026–II–1034. [Online]. Available: http://proceedings.mlr.press/v32/benavoli14.pdf.

[73] A. Benavoli, G. Corani, J. Demšar, and M. Zaffalon, "Time for a change: A tutorial for comparing multiple classifiers through bayesian analysis," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 2653–2688, Jan. 2017, ISSN: 1532-4435. [Online]. Available: https://jmlr.org/papers/v18/16-305.html.

[74] J. M. Pérez, J. C. Giudici, and F. Luque, *pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks*, 2021. arXiv: 2106.09462 [cs.CL].

[75] *RoBERTuito, A pre-trained language model for social media text in Spanish*, https://huggingface.co/pysentimiento/robertuito-base-uncased, [Online; accessed 30-May-2023], 2022.

[76] F.-J. Rodrigo-Ginés, J. Carrillo-de-Albornoz, and L. Plaza, "Hierarchical Modeling for Propaganda Detection: Leveraging Media Bias and Propaganda Detection Datasets," in *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023*, ser. CEUR Workshop Proceedings, vol. 3496, Jaén, Spain: CEUR-WS.org, 2023. [Online]. Available: https://ceur-ws.org/Vol-3496/dipromats-paper7.pdf.

[77] J. A. García-Díaz and R. Valencia-García, "UMUTeam at Dipromats 2023: Propaganda Detection in Spanish and English Combining Linguistic Features with Contextual Sentence Embeddings," in *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023*, ser. CEUR Workshop Proceedings, vol. 3496, Jaén, Spain: CEUR-WS.org, 2023. [Online]. Available: https://ceur-ws.org/Vol-3496/dipromats-paper5.pdf.

[78] V. Ahuir, L. F. Hurtado, F. García-Granada, and E. Sanchis, "ELiRF-VRAIN at DIPROMATS 2023: Cross-lingual Data Augmentation for Propaganda De-

tection," in *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023*, ser. CEUR Workshop Proceedings, vol. 3496, Jaén, Spain: CEUR-WS.org, 2023. [Online]. Available: https://ceur-ws.org/Vol-3496/dipromats-paper6.pdf.

[79] D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, and J. Camacho-collados, "TimeLMs: Diachronic language models from Twitter," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, V. Basile, Z. Kozareva, and S. Stajner, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 251–260. DOI: 10.18653/v1/2022.acl-demo.25. [Online]. Available: https://aclanthology.org/2022.acl-demo.25/.

Appendices

7.A List of propagandist sources considered for *Propitter*'s construction.

Table A1: Propagandist sources of data considered for the construction of *Propitter* (124 sources).

Sources
Activist Mommy/Elizabeth Johnston • Al Bawaba • AliciaFixLuke (BB4SP) • American Lookout • American Principles Project (approject) • American Thinker • AmericaTFP • AntiEmpire • Arab News • Beijing Review • Big League Politics • Bipartisan Report • Blunt Force Truth • Breitbart • Caixin • Caldron Pool • China Daily • China Global Television Network (CGTN) • Christian Action Network (Martin Mawyer) • CNS News • Competitive Enterprise Institute (ceidotorg) • Conservative Daily News • Conservative Free Press • Conservative Patriot (Co-firing-line) • CWforA • Daily Mail • Dan Bongino • DavidBartonWB • Defiant • Deplorable Kel • Discover the Networks (Leftist Networks) • Echo Check (The Other Checker) • Eh Conservative • en_volve • engpravda • Europe-Israel • FactCheckingTR • FAIR (The Federation for American Immigration Reform) • FocusFamily • Freedom First • Frontpage Magazine (David Horowitz) • GatestoneInst • Ghost.Report • GOP.gov (House Republicans) • Granma (Cuba) • Gulf News • HeartlandInst • I Hate the Media • Independent Sentinel • Information Liberation (Info Lib News) • Institute for Historical Review • JD Rucker (NOQ Report) • Judicial Watch • Just The News • Left Action • Lew Rockwell • Life News • Mad World News • meforum • Mehr News Agency • Middle East Media Research Institute (MEMRI) • Mondoweiss • MoonBattery • MoveOn • National Right to Life Committee (NRLC) • News Rescue • News18 • newsblaze • nicolejames • Nordic Monitor • NowThis News • Occupy Democrats • other98 • Patriot Journal • PJ Media • Political Dig • PragerU • PRISource • ProFamOrg • Public Discourse • Raheem Kassam • ReadTheHornNews • Red Voice Media • redicetv • remnantnews • Right Side Broadcasting Network (RSBN) • RT • Rudaw • Ruptly • Russia Insider • Russian News Agency-TASS • scrowder • Sean Hannity • Shafaq News • Sputnik News • StopSocialists • Swarajya • syria_updates • takimag • Tasnim News Agency • Tehran Times • TeleSUR • The Blaze • The Clover Chronicle • The Colorado Herald • The D.C. Clothesline • The Daily Wire • The Federalist (FDRLST) • The Free Telegraph • The National (UAE) • The Political Insider • The RFAngle • The Scoop • The Unz Review • the_majalla • TheBell_News • Tim Brown (Reformed Media/FPPTim) • ToddStarnes.com • Trending Politics • Turning Point USA • vdare • WayneDupree.com • WestJournalism • Women are Human

7.B List of non-propagandist sources considered for *Propitter*'s construction.

Table A2: Non-propagandist sources of data considered for the construction of *Propitter* (120 sources).

Sources
ABC News • ABC11 Eyewitness News • ABS-CBN News • Africa News • Ahram Online • Al Arabiya • Al Jazeera • Al-Masdar News (AMN) • Arizona Daily Star (Tucson Star) • Arutz Sheva (Israel National News) • Atlanta Black Star • Atlanta Jewish Times • Austin American-Statesman • Baltimore Sun • Berkshire Eagle • Boston Globe • Boy Genius Report (BGR) • Business Insider • Calgary Sun • CBS News • Charlotte Observer • Chicago Tribune • Citizens for Legitimate Government (CLG News) • CNN Business • CNN Communications • CNN International • America Conservative Review • Daily Hive • Daily Press • Daily Signal • Deadline Hollywood • Deccan Herald • Edmonton Journal • Euronews • Fox News (foxnews.com) • Fresno Bee • Greensboro News and Record • Haaretz • Honolulu Star-Advertiser • Huffington Post (HuffPost) • Hurriyet Daily News • IPOWER • Japan Times • Jewish Standard • JewishNewsUK • Kansas City Star • Kansasdotcom • KMOV • KOCO News 5 • Korea Herald • KUOW NPR • LA Times (Los Angeles Times) • Lethbridge Herald • Lexington Herald Leader • Middle East Monitor • Montreal Gazette (mtlgazette) • MSN.com (MSN News) • mySA • National Review • New Republic • New York Daily News • Newsweek • Northwest Arkansas Democrat-Gazette • NYJewishWeek • Outside The Beltway • Ozy Media • Politico • PressTV • RajaPetra (Malaysia Today) • Raleigh News and Observer • RedState • Roanoke Times • RTDNews • RTE (Radio Television of Ireland) • Sacramento Bee • Santa Barbara Independent • Saudi Gazette • SFGate • Sky News UK • St. Louis Post-Dispatch • Star Tribune • Stars and Stripes • Tablet Magazine • Taiwan News • Tampa Bay Times • Texas Tribune • The Bangkok Post • The Cipher Brief • The Courier-Mail (Australia) • The Daily Tarheel • The Day (New London) • The Hartford Courant • The Herald (Everett) • The Hill • The Jakarta Post • The Japan News • The Nation • The New Humanitarian • The News International • The Olympian • The Patriot-News (Pennlive.com) • The Providence Journal • The Santa Fe New Mexican • The State Newspaper • The Stream • The Sun • The Tacoma News Tribune • The Week (USA) • Thomson Reuters Foundation • Time Magazine • Times Colonist • Times of India • Times of Israel • Utah Public Radio (UPR) • Vancouver Sun • Washington Post • WGN News • Windsor Star

7.C Bias distribution of tweets in main partitions of *PropitterX*.

Table A3: Distribution of tweets per bias and class in main partitions of *PropitterX*

Bias	**Prop.**	**Non-prop.**	**Total tweets**
EXTREME LEFT	504	0	504
FAR-LEFT	3,736	0	3,736
LEFT	5,121	19,424	24,545
LEFT-CENTER	3,318	146,852	150,170
LEAST BIASED	0	23,741	23,741
RIGHT-CENTER	10,355	53,132	63,487
RIGHT	15,394	20,245	35,639
FAR-RIGHT	12,104	0	12,104
EXTREME RIGHT	35,826	0	35,826
RIGHT-CONSPIRACY/PSEUDOSCIENCE	0	1,191	1,191
CONSPIRACY-PSEUDOSCIENCE	0	1,209	1,209
UNKNOWN	4,308	14,032	18,340
		Total	370,492

7.D Training partitions with proportional sampled emotions in *PropitterX-EMO*.

Table A4: Partition considered to train a classifier with proportional sampled emotions in *PropitterX-EMO*.

	Train	
Emotion	Propaganda	Non-propaganda
Fear	19,200	19,200
Anger	6,000	6,000
Joy	2,700	2,700
Sadness	1,200	1,200
Surprise	600	600
Disgust	300	300
Neutral	0	0