Alma Mater Studiorum · Università di Bologna

DIPARTIMENTO DI INTERPRETAZIONE E TRADUZIONE
Corso di Laurea magistrale in Specialized Translation (classe LM-94)

TESI DI LAUREA
in NATURAL LANGUAGE PROCESSING

On the Automatic Multilingual Detection of Persuasion Techniques in the News: A Natural Language Processing Approach

CANDIDATO:
Luca Giordano

Relatore:
Alberto Barrón Cedeño

Correlatore:
Adriano Ferraresi

Anno Accademico 2023/2024
Secondo Appello

Acknowledgements

Giunto alla fine anche di questo percorso, bisogna pur mettere la ciliegina sulla torta: voglio ringraziare le persone che mi sono state vicine, nei miei momenti alti e bassi, perché senza di voi sarebbe stato tutto più difficile.

I miei primi ringraziamenti vanno alla mia famiglia, che ancora una volta e sempre rappresenta per me un abbraccio caldo e la fonte della mia energia. Grazie a mia madre Marianna, mio padre Luciano, mio fratello Andrea, mia nonna Ida e mio zio Antonio per avermi spronato a dare il meglio di me in questi due anni giorno dopo giorno, per aver creduto in me a partire fin dalla doppia prova che ho dovuto sostenere, prima per entrare in questo corso di laurea, subito dopo per ottenere la mia borsa di ricerca. Mi avete supportato in ogni esame, in ogni conferenza, in ogni pubblicazione. Finché crederete in me avrò la forza di poter fare tutto.

Un altro ringraziamento lo dedico al mio amico Christopher. Anche se lontani già da tempo, ci sento sempre vicini col cuore. Il tuo aiuto in questi due anni è sempre stato tangibile: d'altronde, con quale coraggio pubblicherei un articolo senza prima raccontarne il contenuto a te durante una sfiancante passeggiata in salita nei boschi spersi chissà dove?

Un brindisi lo dedico al mio gruppo "L'oca alcolica" e alle nostre serate in quel di Forlì. Ringrazio quindi Angela, Alessandro, Martina, Sara e Silvia per tutti gli innumerevoli momenti passati insieme, a partire dalle serate con giochi da tavolo alla Mondadori, passando per le feste a ballare e cantare insieme da Volume, Oltremodo, Controsenso, Kindergarten, Jump e Abbey, le feste di compleanno, le birre al Games Bond la sera stanchi fino ad arrivare alle lezioni, lo studio insieme e ai caffè da Gardelli. In particolare grazie ad Angela e Sara le mie compagne di techno, ad Alessandro con cui condivido tante passioni, punti di vista e discorsi sull'università e la vita, a Martina a cui so di poter raccontare tutto e con cui so di potermi aprire (e viceversa!!!) e a Silvia la mia compagna russista fin dal primo giorno.

Infine, ringrazio la mia metà, la mia partner di vita, la mia fidanzata Luisa. Grazie per essermi stata vicina, tu hai visto i momenti più bassi di questi due anni ed eri presente e parte integrante di quelli più alti. Non dimenticherò mai la sensazione della tua mano nella mia, i tuoi occhi nei miei in tutti i momenti in cui ne avevo bisogno. Grazie perché sei sempre capace di tirarmi su e tornare a farmi splendere come solo tu sai fare. Ancora come allora gioisco al pensiero di condividere con te la mia vita, ma ora la sogno, la vedo, la tocco, e la creiamo insieme.

Abstract

This thesis presents a Natural Language Processing (NLP) approach to automatically detect persuasion techniques in multilingual news content. As the dissemination of propaganda becomes increasingly prevalent in digital media, identifying these techniques is critical to promoting information literacy and countering propaganda.

Through a comprehensive literature review, this thesis first explores the landscape of propaganda, introducing the concept and its history, and then dives deeper into computational propaganda and its dissemination and detection, highlighting gaps and challenges in current methodologies, both from text and network analysis perspectives. The collective research efforts known as shared tasks are also discussed in detail.

The core of this work focuses on two experiments, involving the participation to the shared task CheckThat! Lab 2024 at CLEF-2024, that challenges participants to develop models to classify persuasion techniques in news across multiple languages, and an attempt conducted months later to enhance the performance of the first system developed. The dataset provided for training includes news articles in multiple languages on several topics annotated with various persuasion techniques, and multiple models are evaluated to determine their effectiveness in a competition setting with a public leaderboard. The methodology of our team UniBO encompasses data preprocessing, data augmentation, and the development of a sophisticated system for detection of persuasion techniques using state-of-the-art NLP tools.

Results show that, while our proposed model achieves competitive performance in multilingual settings, challenges such as data scarcity, explainability, and model generalization persist. Ethical considerations, including biases in detection algorithms, are also addressed. The thesis concludes by discussing limitations and potential avenues for future research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Media language and news discourse have always attracted the attention of applied linguists and sociolinguists, mainly because of four reasons (Bell, 1995):

1. the media provide an easily accessible source of language data for research and teaching purposes

2. the media are important linguistic institutions, and their language usage reflects and shapes both language use and attitudes in a speech community

3. the ways in which the media use language are interesting linguistically in their own right

4. the media are important social institutions and crucial presenters of culture, politics, and social life, shaping as well as reflecting how these are formed and expressed

Beyond the academic and long-term intellectual benefits it provides, analyzing media language and news discourse can also have some positive, immediate effects and tangible, practical applications.

With the widespread use of the Internet and the rise of algorithmic journalism (Anderson, 2013; Coddington, 2015; Graefe, 2016; Thurman, 2018; Túñez-López et al., 2020), characterized by huge amounts of data, the application of algorithms in all phases of the journalistic process (selection, production, distribution and consumption) and by a high degree of interactivity and direct communication between news producers and consumers, the latter are exposed more than ever before to manipulative, propagandistic, and deceptive content. As a result, major public events and debates on important topics can be significantly influenced. As put by Da San Martino et al. (2021), “The issue became of general concern in 2016, a year marked by micro-targeted online disinformation and misinformation at an unprecedented scale, primarily in connection to Brexit and the US Presidential campaign; then, in 2020, the COVID-19 pandemic also gave rise to the first global infodemic. Spreading disinformation disguised as news created the illusion that the information was reliable, and thus people tended to lower their natural barrier of critical thinking compared to when information came from different types of sources”. However, manually fighting online, large-scale, world-wide disinformation and influence/propaganda campaigns, employing human professionals with domain expertise who perform careful media analysis across countries and multiple languages and contexts, is a rather slow and challenging task in today’s media landscape characterized by massive information production. Manual analysis is not only impractical and too slow on a large scale, but can also be inconsistent. Furthermore, the multilingual and cross-cultural variables add complexity, given that factors such as linguistic variation, cultural nuances and subtleties of rhetorical strategies of new contexts can hinder anti-disinformation and anti-persuasion campaigns that proved to be successful in the past.

This led to an increasing demand for efficient, consistent, and automated tools that help experts analyze the news ecosystem, detect manipulation attempts, and aid in studying how events, global issues, and policies are portrayed by the media in different countries and languages. An important aspect of the problem that is often largely ignored is the mechanism through which disinformation is conveyed, that is using propaganda techniques. These include specific rhetorical and psychological techniques, ranging from leveraging on emotions (such as using loaded language, flag waving, appeal to authority, slogans, and clichés) to using logical fallacies (such as straw man, red herring, black-and-white fallacy, and whataboutism). Comprehensive definitions and different taxonomies of persuasion techniques can be found in Section∼2.1. There has been a growing interest of the Natural Language Processing (NLP) community in trying to detect the use of specific propaganda techniques, as well as the specific span of each instance within the text. This work shall be contextualized within this research area.

This thesis introduces the reader to the concept of propaganda and persuasion techniques, with a focus on computational propaganda in news and in multilingual contexts, providing an in-depth, thorough overview of the scientific literature published on this subject and highlighting gaps, challenges, limitations, and ethical considerations. The main objective of this work is developing a system for the automatic, multilingual detection of persuasion techniques in the news that performs better than state-of-the-art approaches. Two versions of the proposed system were developed: the first version was developed in the context of the participation to the shared task *CheckThat! Lab 2024 Task 3 on Persuasion Techniques* (Piskorski et al., 2024) at CLEF 2024. The task consisted in detecting 23 persuasion techniques at the fragment-level in online media and covered highly-debated topics e.g., the Isreali–Palestian conflict, the Russia–Ukraine war, climate change, COVID-19, and abortion. The second version was developed months later, as an attempt to enhance the performance of the first version.

This thesis builds upon the foundational work presented in a research paper I co-authored, titled:

> Gajo, P., Giordano, L., & Barrón-Cedeño, A. (2024). UniBO at CheckThat! 2024: Multi-lingual and Multi-label Persuasion Technique Detection in News with Data Augmentation and Sequence-Token Classifiers. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)* (pp. 426-434).

Expanding on the insights and methodologies developed in that study, this work contributes to the endeavor of developing an organized understanding of what (computational) propaganda is, how it is conveyed in the media today, and how to identify it. The results of the two experiments contribute to this field by highlighting the major challenges of the task and the limitations of contemporary approaches, but also what might be potential solutions or hints toward better performing systems.

The structure of this thesis is the following:

Chapter 2 contains a thorough literature review, first introducing the concept and the history of propaganda and persuasion techniques, some of the existing taxonomies, and the differences between propaganda and disinformation. Then, the chapter delves deeper into the literature, introducing the concept of computational propaganda and analyzing the published works both on its dissemination and detection (from the text analysis and the network analysis perspectives). Follow discussions on the history of the shared tasks organized on this problem.

Chapter 3 defines the task in details, with an overview of the problem, the

data available and evaluation metrics.
Follows **Chapter 4**, where the experiments are thoroughly explained and their results are presented and discussed.
The closing chapter are the **Conclusions** that can be drawn from this work, including a summary of findings, main contributions, limitations, ethical considerations and hints for future work.

Chapter 2

Literature Review

2.1 Propaganda and Persuasion Techniques

According to Jowett and O'Donnell (2018), the modern Latin term 'propaganda', in its most neutral sense, means to disseminate, spread, or promote particular ideas. It was coined in 1622, when the Vatican established the new administrative body *Sacra Congregatio de Propaganda Fide*, and originally referred to the propagation of the Catholic faith in the New World carried out by this institution. Since the term was strongly associated from the very beginning with the conversion to Christianity and the opposition to Protestantism, it soon lost its neutrality, and subsequent usage has conferred the term a pejorative meaning. Indeed, from the end of the 18th century the term began being used also to refer to propaganda in secular activities, and by the mid-19th century it was used in the political sphere (Diggs-Brown, 2011). According to Jowett and O'Donnell (2018), words frequently used in the context of propaganda that testify to the negative connotation are *lies, distortion, deceit, manipulation, mind control, psychological warfare, brainwashing,* and *palaver.*

In 1937 a group of journalists, educators, and business leaders worked to raise awareness about the role of propaganda in contemporary culture, and thus an independent, US-based organization that functioned as a proto-media literacy group of its time was born, called the Institute for Propaganda Analysis (IPA) (Hobbs and McGee, 2014). The main aim of the IPA was educational in its nature, i.e. to help the public detect, recognize and analyze propaganda, and more in general to promote critical thinking and analysis, essential skills to exercise against the emergent mass media (e.g., radio, film, newspapers). The IPA produced many influential papers and instructional materials, among which the seminal work titled "How to Detect Propaganda" by Miller (1939), one of the founders of the IPA. In this work the author defined propaganda as "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". As Bolsover and Howard (2017) note, this definition entails two key elements: i) trying to influence opinion or behavior, and ii) doing so on purpose. Other key elements are the presence of a sizable target audience, the representation of a specific group's agenda and the use of faulty reasoning and/or emotional appeals.

This work is best known for categorizing propaganda into seven rhetorical devices, later called persuasion techniques. Rooted in ancient classical rhetoric, they remain widely accepted today (Jowett and O'Donnell, 2018, p.237-238): *name calling, glittering generalities, transfer, testimonial, plain folks, card stacking* and *bandwagon* (Table∼2.1). Miller states that "we are fooled by propaganda chiefly because we don't recognize it when we see it", but identifying it becomes possible "if we are familiar with the seven common propaganda devices" (Miller, 1939, p.27). He also stresses that these devices appeal to our emotions rather than our reason and critical thinking, and are thus impossible to counter by means of logical reasoning alone: the public must learn about them first and then recognize them as such. Persuasion techniques can be understood as the mechanism through which propaganda and disinformation are conveyed (Da San Martino et al., 2021).

Miller's taxonomy, although widely accepted, is not the only one: the set of propaganda techniques differs between scholars and sources. For example, Weston (2018) lists at least 24 techniques, the crowdsourced list on Wikipedia includes about 70 techniques[1], and Conserva (2003) goes as far as listing 89 techniques. However, the larger sets of techniques are mostly sub-types of the general set proposed by Miller (1939) (Da San Martino et al., 2021). For example, the technique *half-truth* listed on Wikipedia, defined as "a deceptive statement that includes some element of truth [...]" can be considered as a sub-type of the technique *card stacking* proposed in Miller (1939).

Commenting on Miller's 7 propaganda devices, Hobbs and McGee (2014)

[1] https://en.wikipedia.org/wiki/Propaganda_techniques

note that “it’s important to note that the list of rhetorical devices is explicitly presented as knowledge needed to avoid being victimized by a presumably powerful and manipulative persuader [...] The repeated use of the word ‘trick’ in the formulation of the seven propaganda devices suggests that the rhetorical tools themselves are somehow inherently immoral or unethical practices of communication. Given the rise of Fascism, this approach is not surprising but it does seem inconsistent with earlier articulations of propaganda as potentially either ‘good’ or ‘bad’ depending on the motives of the communicator” (Hobbs and McGee, 2014).

Furthermore, Da San Martino et al. (2021) describe the difference between propaganda and disinformation: the main difference lies in the truth value of the information and its goal. Researchers working on disinformation both within and outside the NLP community today mostly adopt the official definition formulated by the European Union, which states that “disinformation is understood as verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm” (EUCommission, 2022). Contrarily, propaganda, despite its acquired dominant negative connotation, can include claims that are either true or false, and its intended objectives can be either harmful or harmless, with the latter factor being subjective depending on the observer.

An especially productive line of inquiry within propaganda studies involves analyzing propaganda in news media, as it reveals how information can be subtly or overtly manipulated to shape public perception and opinion. By examining the language, framing, and selection of stories, researchers can identify patterns that serve specific ideological or political agendas. This analysis not only uncovers the mechanisms of influence but also empowers audiences to become more critical consumers of media, recognizing biases and questioning narratives that may otherwise go unchallenged. A clear example of the impact of propaganda and disinformation in news media is the weaponization of the news cycle during both the 2016 US Presidential elections and the Brexit referendum, which led to the general public being concerned about the dangers of the proliferation of fake news (Howard and Kollany, 2016; Faris et al., 2017; Lazer et al., 2018; Vosoughi et al., 2018; Bovet and Makse, 2019). According to Yu et al. (2021), there are two main take-aways from the 2016 news weaponization: “First, disinformation disguised as news created the illusion that the information is reliable, and thus people tended to lower their barrier of doubt compared to when information

Name	**Definition**	**Example**
1. Name Calling	The propagandist appeals to our hate and fear by giving "bad names" to those individuals, nations, races, policies, practices, beliefs and ideas which he would have us condemn and reject.	Father Coughlin calls President Franklin D. Roosevelt "a liar."
2. Glittering Generalities	The propagandist identifies his program with virtue by use of "virtue words" such as truth, freedom, honor, social justice, democracy.	"What America needs," says Roosevelt, "is economic security for all."
3. Transfer	The propagandist carries over the authority, sanction, and prestige of something we respect and revere to something he would have us accept.	Roosevelt made a symbol of the horse and buggy when he spoke of an anti-New Deal Supreme Court decision.
4. Testimonial	The propagandist secures statements or letters from prominent people with the expectation that the crowd will follow the leader.	If large numbers of individuals can be seen voting for Roosevelt or for Landon, it is likely to cause many additional votes for them.
5. Plain Folks	The propagandist attempts to win our confidence by appearing to be common people like ourselves.	It is proverbial that political candidates always kiss babies.
6. Card Stacking	The propagandist tells us only part of the truth, uses under-emphasis and over-emphasis to dodge issues and evade facts, and confuses and diverts those in quest of the truth.	In 1936, with unemployment still the serious issue in America, the Republican propagandists blame the Democrats for not ending it.
7. Band Wagon	The propagandist attempts to make us follow the crowd, leading to mass acceptance of the political program.	Everybody's doing it.

Table 2.1: The 7 propaganda devices proposed by Miller (1939). Examples borrowed from Hobbs and McGee (2014).

came from other types of sources. Second, the rise of citizen journalism led to the proliferation of various online media, and the veracity of information became an issue".

Given the huge amount of information produced everyday online that needs to be fact-checked and analyzed for propaganda or disinformation, there is an increasing demand for automated tools that help experts analyze the news ecosystem, detect manipulation attempts, and aid in studying how events, global issues, and policies are portrayed by the media in different countries and languages. This led to a growing interest of the Natural Language Processing (NLP) community in trying to detect the presence of propaganda and specific persuasion techniques in news media.

2.2 The Dissemination of Computational Propaganda

As the proliferation of online content continues to rise, so does the strategic use of language to influence public opinion, often in subtle or deceptive ways. As noted by Bolsover and Howard (2017), propaganda campaigns had traditionally been a monopoly of state actors, whereas nowadays, due to the rise of algorithmic journalism and social media (Anderson, 2013; Coddington, 2015; Graefe, 2016; Thurman, 2018; Túñez-López et al., 2020), they are within reach for various groups and even for individuals. Nonetheless, propaganda campaigns still often rely on large, coordinated efforts to spread messages at scale and make an impact, and this coordination is usually achieved by leveraging computational, technical means such as bots (groups of fully automated accounts) and cyborgs (partially automated) (Chu et al., 2012; Zhang et al., 2016; Da San Martino et al., 2021; Musser, 2023). The type of propaganda "created or disseminated using computational (technical) means" is called computational propaganda (Bolsover and Howard, 2017).

The study of computational propaganda is ever more relevant in the era of Large Language Models (LLMs), since several scholars have speculated that they may be used by malicious actors to generate divisive, misleading, or false information for the purpose of social manipulation (Buchanan et al., 2021; Bagdasaryan and Shmatikov, 2022; Kreps et al., 2022; Patel and Sattler, 2023), and some of the organizations releasing such new technologies have explicitly acknowledged this as a misuse risk (Radford et al., 2019; Weidinger et al., 2021). With the aim of investigating whether there is any economic benefit for malicious actors in using technologies for the automatic generation of text, in particular LLMs, and whether monitoring controls and policies on API-accessible models can have any impact on their misuse, Musser (2023) conducted a cost analysis and constructed a model that aims at approximating the real costs that propagandists bear for automatic social media content generation at scale and their coordinated influence operations. The experiments show "that LLMs need only produce usable outputs with relatively low reliability (roughly 25%) to offer cost savings to propagandists, that the potential reduction in content generation costs can be quite high (up to 70% for a highly reliable model), and that monitoring capabilities have sharply limited cost imposition effects when alternative open source models are available" (Musser, 2023). Furthermore, the findings suggest that "nation-states, even those conducting many large-scale influence operations per year, are unlikely to benefit economically from training custom LLMs specifically for use in influence operations". This suggests that malicious, propagandistic actors today can easily use any of the off-the-shelf LLMs available, private or open-source depending on the degree of monitoring on the private ones and on the performance of the open-source ones, and conduct fully or semi-automatic, large-scale influence operations at a low cost relatively to an influence operation conducted by humans alone.

2.3 The Detection of Computational Propaganda

Although the computational, technical aspect of propaganda today is a key element that cannot and should not be underestimated, Bolsover and Howard (2017) argue that "viewing computational propaganda only from a technical perspective, as a set of variables, models, codes, and algorithms, plays into the hands of those who create it, the platforms that serve it, and the firms that profit from it. The very act of making something technical and impartial makes it seem inevitable and unbiased. This undermines the opportunities to argue for change in the social value and meaning of this content and the structures in which it exists". If on the one hand it is reasonable to argue for the need of research on detecting, exposing, and countering propaganda, on the other hand, specifically in the context of the political domain, Bolsover and Howard (2017) note that "prediction, models, and technical solutions should not be the primary goal of political big-data research", but rather "variables and models are important for what they tell us about underlying social phenomenon": "Too many big-data studies report only the predictive power of their models. However, prediction is not the goal; understanding is the goal". Thus, as articulated by Yu et al. (2021), "interpretability is indispensable if propaganda detection systems are to be trusted and accepted by the users. [...] even if a model can correctly predict which news is propagandistic, if it fails to explain the reason for that, people are more likely to reject the results and to stick to what they want to believe".

The automatic detection of propaganda online has been approached from two main different perspectives: the text analysis perspective and the network analysis perspective. Da San Martino et al. (2021) points out that there is a disconnection between the NLP and Network Analysis communities, and therefore recommends exploring hybrid approaches that may lead to outperforming the state-of-the-art. An example of a hybrid approach is the experiment conducted by Hristakieva et al. (2022), where the authors explored the interplay between propaganda and coordination in the 2019 UK electoral debate on Twitter. For the network analysis to measure the extent of coordination, the authors followed the approach of Nizzoli et al. (2021), while for the text analysis to measure the presence of propaganda they used the Proppy classifier created by Barrón-Cedeño et al. (2019).

2.3.1 Text Analysis Perspective - Datasets and Models

Research on automatic persuasion technique detection in news and on social media overlaps to a large extent with work on automatic propaganda detection in news and on social media (Rashkin et al., 2017; Barrón-Cedeño et al., 2019; Da San Martino et al., 2019b, 2020b; Wang et al., 2020; Li et al., 2022; Sprenkamp et al., 2023; Maarouf et al., 2023; Salman et al., 2023; Sajwani et al., 2024), and both have a short history due to the lack of suitable annotated datasets for training supervised models (Da San Martino et al., 2021). Regardless of type of content, early research on propaganda detection focused exclusively on document-level analysis, ignoring the fine-grained aspects of the task.

Rashkin et al. (2017) created the TSHP-17 corpus, a collection of 22,580 news articles annotated in a distant supervised manner (i.e. assigning the label of a news outlet as judged by media company *US News & World Report*[2] to all articles gathered from that news outlet) at the document-level with four balanced classes: trusted, satire, hoax, and propaganda. However, as can be deduced from the results obtained in the experiment, further verified for reproducibility by Barrón-Cedeño et al. (2019), and mentioned by Da San Martino et al. (2021), the predictive model trained on this data (logistic regression with n-gram representation) failed to generalize, performing well only on articles from sources that the system was trained on and underperforming when evaluated on articles from unseen news sources.

Barrón-Cedeño et al. (2019) created the QProp corpus, an imbalanced collection of 51,294 news articles annotated with distant supervision from fact-checker *Media Bias/Fact Check*[3] at the document-level with two labels (propaganda vs non-propaganda) and trained different models (e.g., logistic regression and SVMs) on this data and on the TSHP-17 corpus to predict the two classes, including linguistic features such as writing style and readability indices. Their findings confirmed that using distant supervision might introduce bias in the model and lead to predict the source of the article, rather than to discriminate propaganda from non-propaganda, independently from the news source.

Wang et al. (2020) presented an approach to leverage cross-domain learning in propaganda detection across different domains, i.e. news, tweets and political speeches (Figure∼2.1). The authors proposed a set of linguistic features (although the choice of the authors to consider TF-IDF and N-gram semantic representations as linguistic features is debatable) and built various classifiers trained on five datasets from the three domains to test cross-domain learning capabilities. The experimental results show that the best cross-domain performance is obtained when training on news and inferencing on speeches or tweets. However, cross-domain learning in propaganda detection still proves to be a challenging task. Furthermore, there is no feature set among the ones tested that can be claimed to be the absolute best, suggesting that different datasets exhibit different kinds of linguistic cues. Finally, the authors note that when excluding proper nouns, a notable drop in performance can be

[2] US News & World Report

[3] Media Bias/Fact Check

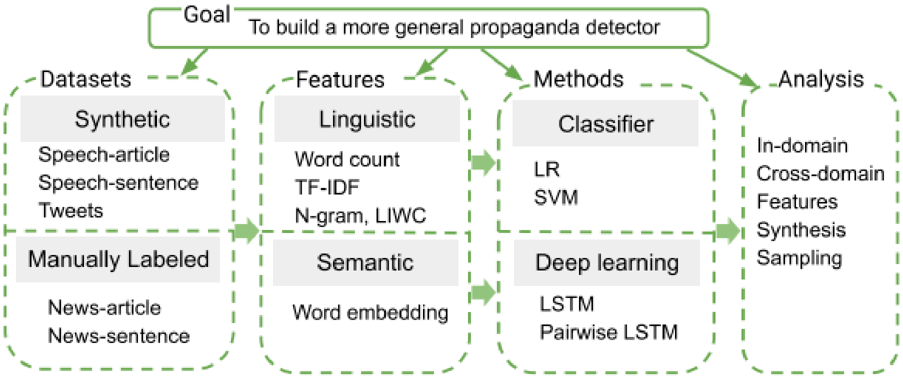


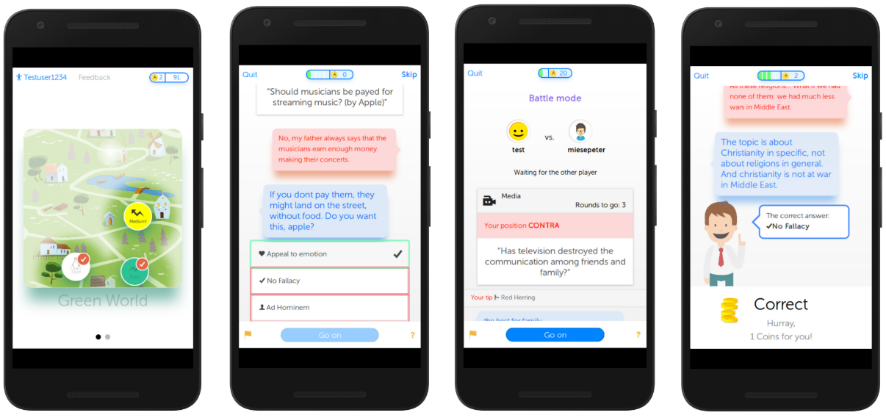
Figure 2.1: Framework for propaganda detection and analysis by Wang et al. (2020).

observed, highlighting the relevance of the entities to which news, tweets or speeches refer.

Maarouf et al. (2023) released the HQP dataset, a collection of 30,000 English tweets on the Russo-Ukrainian war manually annotated by crowdsourced annotators at document-level with two labels (propaganda vs non-propaganda). The authors then ran several classification experiments with the aim of highlighting any differences in the performance of models trained on datasets annotated with distant supervision and datasets annotated manually. The experimental results show that then state-of-the-art language models failed in detecting on-line propaganda when trained with weak labels (i.e. distant supervision), while they could accurately detect it when trained with high-quality manual annotations. The authors claimed a 44% improvement in the performance of their model trained on the HQP dataset with manual annotations in comparison to a version of their dataset with weak labels.

An alternative line of research has focused on detecting the use of specific propaganda techniques in text (Habernal et al., 2017, 2018; Da San Martino et al., 2019b, 2020b; Yu et al., 2021; Li et al., 2022; Sprenkamp et al., 2023; Salman et al., 2023; Sajwani et al., 2024).

Habernal et al. (2017, 2018), specifically focusing on computational argumentation rather than strictly on propaganda, created a corpus with 1.3k arguments annotated with five logical fallacies, that nonetheless are related to propaganda techniques. The authors also created *Argotario*, a multilingual, open-source, web-based application with strong research-oriented



"Should musicians be payed for streaming music? (by Apple)"

No, my father always says that the musicians earn enough money making their concerts.

If you dont pay them, they might land on the street, without food. Do you want this, apple?

Waiting for the other player

Your position CONTRA

"Has television destroyed the communication among friends and family?"

none of them: we had much less wars in Middle East.

The topic is about Christianity in specific, not about religions in general. And christianity is not at war in Middle East.

The correct answer: ✓No Fallacy

Hurray, 1 Coins for you!

(a) A single *world* with the two first *levels* finished, the third one about to be played, and other to be 'explored'.

(b) The recognize fallacy type *round*.

(c) The *player vs. player* level, now waiting for the opponent's turn.

(d) An example of *hard feedback* in a fallacy recognition round.

Figure 2.2: Screenshots of *Argotario* taken on a smartphone emulator from Habernal et al. (2017).

and educational aspects regarding logical fallacies in argumentative discourse (Figure∼2.2).

A more fine-grained analysis was done by Da San Martino et al. (2019b), who created PTC corpus, a collection of 451 news articles manually annotated by professional annotators at the fragment-level with 18 propaganda techniques. Da San Martino et al. (2019b) defined two tasks, based on annotations from the PTC dataset: i) binary classification: given a sentence in an article, predict whether any of the 18 techniques has been used in it; ii) multi-label multi-class classification and span detection task: given a raw text, identify both the specific text fragments where a propaganda technique is being used as well as the type of technique. The authors trained a multi-granularity gated deep neural network for sentence-level propaganda detection.

Da San Martino et al. (2020b) proposed Prta (Propaganda Persuasion Techniques Analyzer)[4], a publicly accessible online platform that allows users to explore news articles crawled on a regular basis by highlighting the spans in which 18 propaganda techniques occur and to compare them on the basis of their use of propaganda techniques. The system further reports statistics about the use of such techniques, overall and over time, or according to fil-

[4] https://www.tanbih.org/prta

tering criteria specified by the user based on time interval, keywords, and/or political orientation of the source. Moreover, it allows users to analyze any text or URL through a dedicated interface or via an API. The aim of the platform, aligned with the original intentions of the Institute for Propaganda Analysis as reported in Section~2.1, is to make online readers aware of propaganda, promoting media literacy and critical thinking. The platform Prta relies on a supervised multi-granularity gated BERT-based model trained on the PTC corpus (Da San Martino et al., 2019b).

Building up on the endeavor of promoting media literacy and critical thinking, Yu et al. (2021) proposed a model for interpretable propaganda detection, which can explain which sentence in a news article is propagandistic by pointing out the propaganda techniques used, and why the model has predicted it to be propagandistic, which is a challenging feat in the era of black-box neural classifiers, and therefore ever more relevant not only for propaganda detection itself, but for the whole scientific field of explainable artificial intelligence. To this end, the authors devised novel features motivated by human behavior studies (such as the position of the sentence relative to the whole text, the semantic similarity between a given sentence and the title of the article, or the sentiment expressed), quantitatively deduced the relationship between semantic or syntactic features and propaganda techniques, and selected the features that were important for detecting propaganda techniques for a series of classification experiments. The authors claimed that their proposed method can be combined with a pre-trained language model to yield then state-of-the-art results.

Dimitrov et al. (2021a) proposed to extend the persuasion technique detection task to multimodal contexts, i.e. to internet memes. The authors created a corpus of 950 memes about several topics (including vaccines, politics, COVID-19, gender equality and more) gathered from Facebook and manually annotated it with 22 propaganda techniques, which can appear in the text, in the image, or in both. The annotation procedure conducted suggests that both modalities are essential for the task. The authors also experimented with several state-of-the-art textual, visual, and multimodal models. The experimental results show that i) the visual-only model (ResNet-152) performs worse than text-only models (fastText and BERT), ii) the best multimodal fusion model is early fusion BERT + ResNet-152, and iii) the best results are obtained with native multimodal models (ViLBERT CC and Visual-BERT COCO).

Li et al. (2022) presented their approach for detecting propaganda techniques

Technique	**Definition**
Name calling	attack an object/subject of the propaganda with an insulting label
Repetition	repeat the same message over and over
Slogans	use a brief and memorable phrase
Appeal to fear	support an idea by instilling fear against other alternatives
Doubt	questioning the credibility of someone/something
Exaggeration/minimiz.	exaggerate or minimize something
Flag-Waving	appeal to patriotism or identity
Loaded Language	appeal to emotions or stereotypes
Reduction ad hitlerum	disapprove an idea suggesting it is popular with groups hated by the audience
Bandwagon	appeal to the popularity of an idea
Casual oversimplification	assume a simple cause for a complex event
Obfuscation, intentional vagueness	use deliberately unclear and obscure expressions to confuse the audience
Appeal to authority	use authority's support as evidence
Black & white fallacy	present only two options among many
Thought terminating clichés	phrases that discourage critical thought and meaningful discussions
Red herring	introduce irrelevant material to distract
Straw men	refute argument that was not presented
Whataboutism	charging an opponent with hypocrisy

Table 2.2: List of the 18 propaganda techniques of the PTC corpus from Da San Martino et al. (2019b) and their definitions.

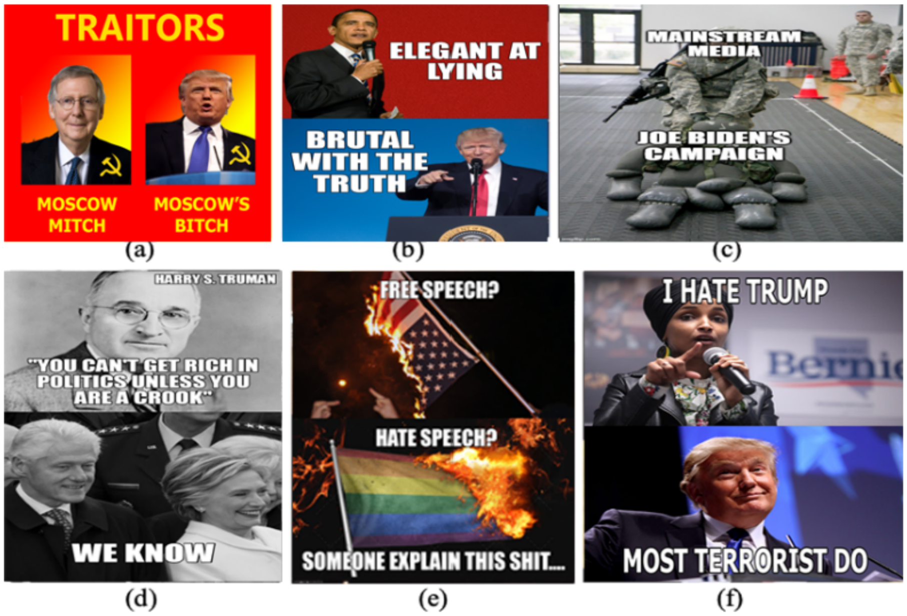


Figure 2.3: Examples of memes from the dataset by Dimitrov et al. (2021a).

in news articles, focusing on two main tasks: Span Identification and Technique Classification. Utilising a BERT model, the authors enhanced their system with an over-sampling strategy and easy data augmentation (EDA) techniques (i.e. synonym replacement and random swap), and introduced a sentence-level feature concatenation method. Their experiments, conducted on a dataset of approximately 550 news articles labeled with 14 propaganda techniques at the fragment-level from the SemEval-2020 task 11 on "Detection of Propaganda Techniques in News Articles" (Da San Martino et al., 2020a), showed state-of-the-art performance in identifying and classifying the techniques.

Sprenkamp et al. (2023) investigated the effectiveness of Large Language Models such as GPT-3 and GPT-4 for propaganda detection. The authors conducted several multi-label classification experiments with five settings with GPT-3 and GPT-4 using the same dataset of Li et al. (2022). To define the five experimental settings, the authors tested various prompt engineering and fine-tuning strategies: for GPT-4, they tested both a 'base' prompt (direct instruction) and a 'chain of thought' prompt, asking the model to engage in a reasoning process about the predicted labels; for GPT-3, they fine-tuned it with the given dataset and then prompted it with the ‘base’ and ‘chain of thought’ prompts. Additionally, the authors fine-tuned another GPT-3 model, this time without including any instructional prompt for the task neither at training nor at inference time. In all settings, the prompts are ‘few-shot’, giving a single example for each propaganda technique within the prompt. The experimental results were compared with the then state-of-the-art approach using RoBERTa (Liu et al., 2019), and claim that GPT-4 achieves comparable performance while significantly reducing the need for labeled training data and resources otherwise needed for supervised learning methods.

Salman et al. (2023) brought attention to the fact that most work on propaganda detection has focused exclusively on high-resource languages such as English, while little effort has been made to detect propaganda for lower-resource languages. The authors argue that it is common to find a mix of multiple languages in a single social media content, a phenomenon known as code-switching (Scotton, 1982; Tay, 1989; Nilep, 2006). Code-switching combines different languages within the same text, which poses a challenge for automatic systems. Considering this premise, they proposed a novel task of detecting propaganda techniques in code-switched text. To support this task, the authors created a corpus of 1,030 texts from Twitter, Facebook, Instagram, and Youtube, code-switching between English and Roman Urdu, and manually annotated it with 20 propaganda techniques at the fragment level. They performed a number of experiments contrasting different experimental setups, including monolingual, multilingual, crosslingual models, and Large Language Models. The findings show that it is important to model the multi-linguality directly rather than using translation as well as to use the right fine-tuning strategy.

Building up on the work of Da San Martino et al. (2020b) and their system Prta, Sajwani et al. (2024) created FRAPPE (Framing, Persuasion, and Propaganda Explorer)[5], a publicly accessible online news analyzer platform that allows the user to unveil the intricate linguistic techniques used to shape readers’ opinions and emotions in news articles. The system allows users not only to analyze individual articles for their genre, framings, and use of persuasion techniques, but also to draw comparisons between the strategies of persuasion and framing adopted by a diverse pool of news outlets and countries across multiple languages for different topics, thus providing a com-

[5] https://frappe.streamlit.app/

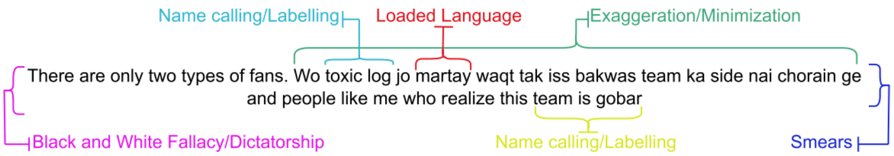

Name calling/Labelling
Loaded Language
Exaggeration/Minimization
There are only two types of fans. Wo toxic log jo martay waqt tak iss bakwas team ka side nai chorain ge and people like me who realize this team is gobar
Black and White Fallacy/Dictatorship
Name calling/Labelling
Smears

Figure 2.4: Example of code-switched text annotated at the fragment level from Salman et al. (2023).

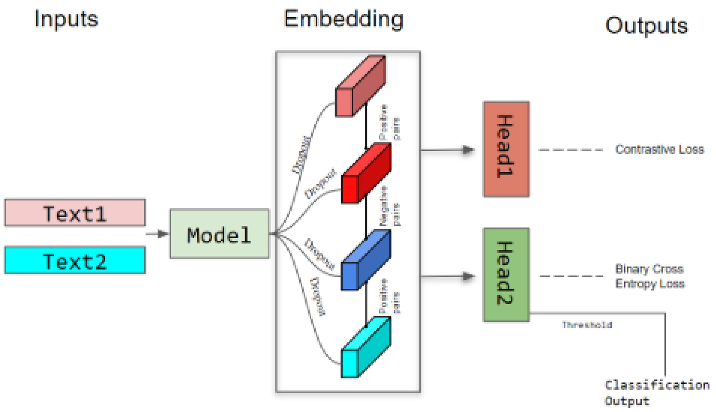


Figure 2.5: Architecture of FRAPPE for framing and propaganda from Sajwani et al. (2024).

prehensive understanding of how information is presented and manipulated. Underlying the platform there are three models trained on the same dataset but focusing on one of its different sets of annotations with respect to the three different functions of the platform. The dataset comes from the shared task SemEval-2023 task 3 on “Detecting the Genre, the Framing, and the Persuasion Techniques in Online News in a Multi-Lingual Setup” (Piskorski et al., 2023c) and consists of 1,612 articles covering news on current topics of public interest in six European languages (English, French, German, Italian, Polish, and Russian), with more than 37k annotated spans. Each news article was annotated for genre, framings, and persuasion techniques.

According to Da San Martino et al. (2021), the main takeaway from the text analysis perspective is that there is a lack of suitable datasets for document-level propaganda detection. Using distant supervision as a substitute is problematic because i) it leads to modeling news sources rather than detecting propaganda, and ii) it is based on the wrong assumption that all articles from a given source would be either propaganda or non-propaganda. Therefore, due to this lack of suitable datasets for document-level propaganda detection and the need for a more fine-grained analysis, the community shifted the focus on detecting the use of specific propaganda techniques at the fragment-level, since they are well-defined in the literature, they are the very device through which propaganda is conveyed, and detecting them allows to develop explainable models. Models capable of pointing out to the user the exact occurrence of persuasion techniques in a news article is a great step toward acceptance of the model output by the user and, from a broader point of view, toward public education on matters of propaganda.

2.3.2 Network Analysis Perspective - Datasets and Models

As shown by Musser (2023), a necessary condition to detect propagandistic campaigns online implies detecting large-scale, malicious coordination, and therefore an analysis that goes beyond individual texts and their merely linguistic characteristics is needed. An example is the identification of the social media users that contributed to injecting and spreading propaganda within a network.

As reported by Da San Martino et al. (2021), “Early approaches for detecting malicious coordination were based on classifying individual nodes in a network as either malicious or legitimate. Then, clusters of malicious nodes were considered to be acting in coordination. In other words, the concept of coordination was not embedded within the models, but it was added “a posteriori”. The vast majority of these approaches are based on supervised machine learning and each account under investigation was analyzed in isolation. That is, given a group of accounts to analyze, the supervised technique was separately applied to each account of the group, that in turn received a label assigned by the detector. The key assumption of this body of work is that each malicious account has features that make it clearly distinguishable from legitimate ones”. Examples of such features are profile characteristics, social network structure, type of content produced (including sentiment expressed) and temporal features. The most widely known example of such approaches is Botometer (Yang et al., 2019), a social bot detection system that simultaneously analyzes multiple dimensions of suspicious accounts for spotting bots. It leverages more than 1,200 features for social media account, evaluating profile characteristics, social network structure, produced content, and temporal features.

However, these early approaches had several limitations, exacerbated by new, evolving threats. First of all, since they were all supervised detectors, they relied on the existence of ground truth training datasets, whereas, in most cases, a real ground truth for bot detection is lacking (Grimme et al., 2017) and labels are manually assigned by humans, who have been proven to suffer from several annotation biases and to fail at spotting sophisticated bots (Cresci et al., 2017). Moreover, it has been demonstrated that malicious accounts "evolve" (i.e. they change their characteristics and behaviors) in an effort to evade detection by established systems, such as Botometer (Cresci et al., 2017). Furthermore, nowadays more advanced and easily accessible technological means such as LLMs and image-video generators exacerbated the problem, enabling propagandists and malicious actors in general to easily generate credible textual and multi-medial content, thus increasing their capabilities of impersonating real people and escape detection.

The limitations of the early approaches led to a new research trend whose primary characteristic is to target groups of accounts as a whole, rather than focusing on individual accounts. Coordination is considered a key feature to analyze, and it is modeled within the detectors themselves. Furthermore, the focus shifted from extensive feature engineering to learning effective, native feature representations for the task and to designing brand-new and customized, task-specific algorithms instead of leveraging general-purpose classification algorithms such as decision trees, random forests and SVMs. Many modern detectors are also unsupervised or semisupervised (e.g., Liu et al. (2017); Chetan et al. (2019)). Other techniques adopted unsupervised approaches specifically for spotting anomalous patterns in the temporal tweeting and retweeting behaviors of groups of accounts (Chavoshi et al., 2016; Mazza et al., 2019). As put by Da San Martino et al. (2021), "The rationale behind such approaches is based on evidence suggesting that human-driven and legitimate behaviors are intrinsically more heterogeneous than automated and inauthentic ones [Cresci et al., 2020]. Consequently, a large cluster of accounts with highly similar behavior might serve as a red flag for coordinated inauthentic behavior".

According to Da San Martino et al. (2021), there are three main takeaways from the network analysis perspective: i) New detectors for coordinated behavior are often developed only after the behavior has been observed, giving malicious actors a significant window of opportunity to exploit online platforms before countermeasures are in place, ii) Most machine learning algorithms used for this task operate under the assumption of a frozen and unchanging context, but this assumption is frequently violated. Malicious accounts evolve over time, and adversaries actively attempt to deceive detectors by changing their behavior and characteristics, and iii) Adversarial machine learning, which accounts for the presence of adversaries by design, could mitigate these issues, since tasks related to detecting online deception and manipulation are intrinsically adversarial.

2.4 Shared Tasks

The growing interest of the NLP community in trying to detect the use of propaganda and its persuasion techniques is mainly expressed by the organization of several shared tasks. The Special Interest Group for Building Educational Applications of the Association for Computational Linguistics defines shared tasks as "collaborative efforts in which researchers and practitioners come together to solve a common problem using shared data and evaluation measures. They promote competition, collaboration, and progress in research, and have become an important part of many academic and industrial communities"[6]. Elstner et al. (2023) says that "in computer science, a shared task is a friendly research competition in which solutions to a given challenging research problem, formulated as a computational task, are developed by several independent teams and then comparatively evaluated. Typical results of such a "shared experiment" are an overview of the effectiveness and efficiency of state-of-the-art approaches to solve the task, but also standardized benchmarks, often adopted by the respective community. Participants in shared tasks are usually asked to describe their approaches in a paper. The organizers then publish a technical report on the benchmark, the experimental setup, the participants' solutions and their performance in solving the task". Shared tasks might also be a promising and effective way to better link research and teaching (Healey, 2005; Elstner et al., 2023). Given that shared tasks usually tackle very challenging and difficult problems, and

[6] Shared Tasks - SIGEDU

that automatic propaganda and persuasion technique detection can be considered one of them, many such events have been organized on this task.
The first shared task on this topic has been the *NLP4IF-2019 Task on Fine-Grained Propaganda Detection* (Da San Martino et al., 2019a), organized as part of the NLP4IF workshop at EMNLP-IJCNLP 2019. There were two subtasks. FLC (fragment-level classification) is a task that asks for the identification of propagandist text fragments in a news article and also for the prediction of the specific propaganda technique used in each such fragment (18-way classification task). SLC (sentence-level classification) is a binary classification task asking to detect the sentences that contain propaganda. The participants were provided with the PTC corpus (Da San Martino et al., 2019b) and 47 additional articles were annotated for the task. Although the fragment-level task proved to be much more challenging than the sentence-level task, most systems easily managed to beat the trivial baselines (a simple logistic regression classifier with default parameters with the sentence length as the only input feature for the SLC task and a random span generator and technique classifier for the FLC task) by a sizable margin for both subtasks. Almost all of the systems proposed, including the best ones, were fine-tuned transformer-based models such as BERT (Devlin et al., 2019), but also other classifiers have been used (or ensembles thereof), such as LSTMs (Hochreiter, 1997), neural networks combined with word or character embeddings, CNNs (LeCun et al., 1998), hand-crafted features and also other machine learning algorithms such as logistic regression.
Followed the *SemEval-2020 Task 11 on Detection of Persuasion Techniques in News Articles* (Da San Martino et al., 2020a). The task featured two subtasks. Subtask SI is about Span Identification: given a plain-text document, spot the specific text fragments containing propaganda. Subtask TC is about Technique Classification: given a specific text fragment, in the context of a full document, determine the propaganda technique it uses, choosing from an inventory of 14 possible propaganda techniques. The participants were provided with the PTC corpus (Da San Martino et al., 2019b), and 90 additional articles for testing were annotated for the task. The new corpus, comprehensive of the new articles, was called PTC-SemEval20 Corpus. Just like for the previous shared task, the subtask SI (segment identification) was easier and all systems proposed managed to improve over the trivial baseline (a random span generator that selects the starting character of a span and then its length), whereas the subtask TC (technique classification) proved to be much more challenging, and this time some teams could not improve over

the still trivial baseline (a logistic regression classifier with default parameters with the sentence length as the only input feature). For both subtasks, the best systems used pre-trained Transformers (mainly BERT and RoBERTa (Liu et al., 2019) models) and ensembles along with some form of data augmentation and post-processing. Other classifiers used mainly include LSTMs and CNNs, and the most used feature representations were embeddings and ELMo (Peters et al., 2018), but also linguistic features (such as part-of-speech) and sentiment-related features.

The following year at SemEval a new multimodal task was proposed: *SemEval-2021 Task 6 on Detection of Persuasion Techniques in Texts and Images* (Dimitrov et al., 2021b). The task focused on memes and had three subtasks: i) detecting the techniques in the text, ii) detecting the text spans where the techniques are used, and iii) detecting techniques in the entire meme, i.e. both in the text and in the image. The participants were provided with the same dataset used in Dimitrov et al. (2021a). The evaluation results for the third subtask confirmed the importance of both modalities, the text and the image. Moreover, some teams reported benefits when not just combining the two modalities, e.g., by using early or late fusion, but rather modeling the interaction between them in a joint model. In all three subtasks the best and most frequently used systems for the textual modality were fine-tuned transformer-based models such as BERT and RoBERTa. In subtask three the most common representations for the visual modality were variants of ResNet. For the textual modality, across the subtasks, techniques such as ensembles, data augmentation and post-processing were also used.

The following year another shared task on this topic was organized, this time as part of the WANLP-2022 workshop at EMNLP-2022, and it was called *WANLP-2022 Shared Task on Propaganda Detection in Arabic* (Alam et al., 2022). The main aim of this shared task was to bring the attention of researchers in the area of propaganda and persuasion technique detection to the Arabic language. The task included two subtasks: Subtask 1 asked to identify the set of propaganda techniques used in a tweet, which is a multilabel classification problem, while Subtask 2 asked to detect the propaganda techniques used in a tweet together with the exact span(s) of text in which each propaganda technique appears. The participants were provided with a new dataset of 930 tweets in Arabic randomly sampled from those posted by the top-2 news sources from each Arab country in addition to five international sources that broadcast Arabic news. The dataset was manually annotated with the same persuasion techniques studied in Da San Martino et al. (2019b) and Dimitrov et al. (2021b). For both subtasks, the majority of the systems fine-tuned pre-trained Arabic language models (such as AraBERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2021)), and used standard pre-processing. Some systems used data augmentation, ensemble methods and standard preprocessing.

In 2023 the ever-growing interest of the scientific community is expressed by the organization of three tasks rather than one.

The first was *SemEval-2023 Task 3 on Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup*, organized by Piskorski et al. (2023b). The task focused on news articles in nine languages, six known to the participants upfront (English, French, German, Italian, Polish, and Russian), and three additional ones revealed to the participants at the testing phase (Spanish, Greek, and Georgian). The task featured three subtasks: i) determining the genre of the article (opinion, reporting, or satire), ii) identifying one or more frames used in an article from a pool of 14 generic frames, and ii) identify the persuasion techniques used in each paragraph of the article, using a taxonomy of 23 persuasion techniques. The participants were provided with a new, multilingual dataset of 2,049 news articles published in the period between 2020 and mid-2022, and revolving around various globally discussed topics, including the COVID-19 pandemic, abortion-related legislation, migration, Russo-Ukrainian war, and some local events such as parliamentary elections. Both mainstream media and "alternative" media sources that could potentially spread mis/disinformation were considered. The dataset was annotated for genre (document-level: opinion, reporting or satirical), framing (document-level: one or more frames form a pool of 14 framing dimensions introduced in Card et al. (2015)), and persuasion techniques (fragment-level in each paragraph: one or more techniques from the revised version of the taxonomy introduced in Da San Martino et al. (2019a); Dimitrov et al. (2021b), which now included 23 techniques). For subtask 1, almost all participants used transformers. The scarcity of the annotated data was dealt with either by combining the datasets for all languages, e.g., via multilingual language models or by automatic translation, or by looking for similar datasets in the literature; ensemble methods have also been very popular. For subtask 2, Since the models were all transformer-based, what differentiated the participating systems were once again the pre-processing and the data augmentation techniques. The vast majority of teams trained their systems on all languages and used ensembles. For subtask 3, the big picture is very similar to the previous subtasks: multilingual

transformer models were used by all participants, and what differentiated the approaches was again the pre-processing and data augmentation strategies. The second event organized in 2023 was the follow-up of WANLP-2022, i.e. the *ArAIEval-2023 Task 1 on Persuasion Technique Detection* (Hasanain et al., 2023b), part of the first ArabicNLP 2023 conference co-located with EMNLP-2023. The goal of this task was to identify the persuasion techniques present in a piece of text. It targeted multi-genre content, including tweets and paragraphs from news articles. The task was organized into two subtasks. Subtask 1A: Given a text snippet, identify whether it contains content with any persuasion technique. This is a binary classification task. Subtask 1B: Given a text snippet, identify the propaganda techniques used in it. This is a multilabel classification task. The participants were provided with a dataset that sampled, combined and revised with new annotations two existing datasets: the tweets dataset described in WANLP-2022 and the dataset AraFacts (Ali et al., 2021), that contains claims verified by Arabic fact-checking websites, and each claim is associated with web pages propagating or negating the claim. The majority of the systems fine-tuned pre-trained Arabic language models (mainly AraBERT and MARBERT) and used standard pre-processing. Several systems explored different loss functions, while a handful of systems utilized data augmentation and ensemble methods.
The last event organized in 2023 was the shared task *DIPROMATS-2023 - Automatic Detection and Characterization of Propaganda Techniques from Diplomats of Major Powers* (Moral et al., 2023) at IberLEF-2023. Three tasks were proposed for each of the two languages, English and Spanish. The first one aimed at distinguishing if a tweet has propaganda techniques or not. The second task aimed at classifying the tweet into four clusters of propaganda techniques, whereas the third one focused on a fine-grained categorization of 15 techniques inspired by Da San Martino et al. (2019b) but modified to incorporate other techniques proposed by Johnson-Cartee and Copeland (2004) and Hobbs and McGee (2014). The participants were provided with a new dataset of 12,012 annotated tweets in English and 9,501 tweets in Spanish, posted by authorities of China, Russia, United States and the European Union. The approaches adopted by participants were very diverse. Generally, the best systems incorporated some kind of data augmentation that included contextual information in the message analyzed. Some successful approaches conducted bottom-up strategies that focused on the fine-grained level to resolve the more coarse-grained tasks. As expected, systems achieved worse performances as the complexity of the task increased.

The degree of difficulty also seemed to have an impact on the performance of the systems when dealing with different languages: the more complex the task, the wider the gap between English and Spanish models.
Also in 2024 three events on this research area were organized.
As a follow-up to SemEval-2021, *SemEval-2024 Task 4 on Multilingual Detection of Persuasion Techniques in Memes* was organized (Dimitrov et al., 2024). The task targeted memes in four languages, with the inclusion of three surprise test datasets in Bulgarian, North Macedonian, and Arabic. It encompassed three subtasks: i) identifying whether a meme utilises a persuasion technique; ii) identifying persuasion techniques within the meme's "textual content"; and iii) identifying persuasion techniques across both the textual and visual components of the meme (a multimodal task). The participants were provided with a dataset of 10,000 English memes collected from Facebook on topics such as politics, vaccines, COVID-19, gender equality, and the Russo-Ukrainian War, manually annotated with the persuasion techniques presented in Dimitrov et al. (2021b). However, the test dataset did not only contain English memes, but also Bulgarian, North Macedonian and Arabic memes, making this a multilingual task. In this edition of the shared task, the organizers introduced a hierarchy to allow the assignment of high-level categories in case of high uncertainty when predicting persuasion techniques. The persuasion techniques were grouped in a hierarchy, to be more precise a directed acyclic graph, where the leaves of the hierarchy are the 22 persuasion techniques. Fine-tuning transformer-based architectures was the most dominant approach followed by most teams. The majority of teams participating in Subtask 2 considered both the text and image components of the data, utilizing corresponding transformer models. Several teams designed hierarchical classification techniques, to tackle the hierarchy of labels in Subtask 1 and Subtask 2a. As for the surprise languages, at least a third of the submitting teams used automatic translation to translate the datasets into English.
As a follow-up of ArAIEval-2023, *ArAIEval-2024 Task 1 on Unimodal (Text) Propagandistic Technique Detection* & *Task 2 on Multimodal Propagandistic Memes Classification* was organized (Hasanain et al., 2024b), as part of ArabicNLP-2024 co-located with ACL-2024. The organizers presented two tasks: i) detection of propagandistic textual spans with persuasion techniques identification in tweets and news articles, and ii) distinguishing between propagandistic and non-propagandistic memes. The participants to the first subtask were provided with a new dataset that combined a revised version

of the dataset of tweets used in Alam et al. (2022); Hasanain et al. (2023b), which now included tweets about the events in Gaza, and a revised version of the dataset AraFacts (Ali et al., 2021), enriched with 600,000 news articles collected from over 400 news media outlets, covering 14 different broad topics. The data was manually annotated following the approach adopted in previous studies (Hasanain et al., 2023a, 2024a). The participants to the second subtask were provided with the dataset from Alam et al. (2024), a collection of around 3K memes manually annotated as propagandistic vs not-propagandistic, which were collected from different social media (e.g., Facebook, Twitter, Instagram and Pinterest). Across both tasks, it was observed that fine-tuning transformer models such as AraBERT was at the core of the majority of the participating systems.

As a follow-up of SemEval-2023, the last event organized in 2024 was the *CheckThat! Lab 2024 Task 3 on Persuasion Techniques* (Piskorski et al., 2024) at CLEF-2024. The task consisted in detecting 23 persuasion techniques at the fragment-level in online media. The task covered highly-debated topics in the media, e.g., the Isreali–Palestian conflict, the Russia–Ukraine war, climate change, COVID-19, abortion, and more. For training and development the participants were provided with the same dataset of the SemEval-2023 task 3 (Piskorski et al., 2023b), which covered nine languages: English, German, Georgian, Greek, French, Italian, Polish, Russian, and Spanish. For testing the participants were provided with an entirely new dataset that covers five languages: Arabic, Bulgarian, English, Portuguese, and Slovene. English is the only language for which both training/development and test data existed. The main difference between this task and the former competition organized at SemEval-2023 was that the latter focused on the detection of persuasion techniques at the paragraph level, while this task aimed at developing models to detect and to classify persuasion techniques at the fragment level, which represented an additional challenge. The systems submitted by the participants were compared against a baseline (an XLM-RoBERTa-base token classification model in a zero-shot setting which, for each token, predicts the class with a given probability threshold, and then merges adjacent tokens with the same class in a single span) and a task organizers' system, which used a state-of-the-art transformer-based architecture. The participating systems used the same architecture as the task organizers' system, and leveraged data augmentation techniques. The obtained results compared to the results reported in Piskorski et al. (2023b) confirmed that the detection at the fragment level is a harder task, and leaves space for improvement. The experiments reported in this thesis represent both a description of our system submitted to this competition and an attempt at improving the state-of-the-art performance of the models for this task.

lang	**#docs**	**#chars**	**#spans**	**#ne-par**	**#avg-fr**	**#avg-pt**
train						
English	446	2,431K	7201	9498	3.7	16.1
French	158	737K	5595	2196	3.0	35.4
German	132	581K	4501	1484	4.3	34.1
Italian	227	927K	6027	2552	3.8	26.6
Polish	145	765K	2839	2294	5.0	19.6
Russian	143	590K	3399	1876	2.5	23.8
development						
English	90	403K	1801	3127	5.1	20.0
French	53	222K	1586	610	3.0	29.9
German	45	171K	1236	522	4.6	27.5
Italian	76	287K	1934	882	3.9	25.4
Polish	49	264K	985	800	4.9	20.1
Russian	48	163K	739	515	2.3	15.4
test						
English	54	228K	1775	910	4.7	32.9
French	50	181K	1681	510	3.3	33.6
German	50	259K	1904	790	5.7	38.1
Italian	61	245K	2351	953	3.8	38.5
Polish	47	349K	1491	1006	5.9	31.7
Russian	72	161K	944	601	1.2	13.1
Georgian	29	46K	218	161	1.7	14.7
Greek	64	248K	691	947	2.9	10.1
Spanish	30	109K	546	330	2.3	18.2

Table 2.3: Statistics about the training, the development, and the test data from the shared task SemEval-2023 (Piskorski et al., 2023b).

Category	**Description**	**Techniques**
Justification	an argument made of two parts: a statement and a justification	Appeal to Authority, Appeal to Popularity, Appeal to values, Appeal to fear/prejudice, Flag Waving
Simplification	a statement is made that excessively simplify a problem, usually regarding the cause, the consequence or the existence of choices	Causal oversimplification, False dilemma or no choice, Consequential oversimplification
Distraction	a statement is made that changes the focus away from the main topic or argument	Straw man, Red herring, Whataboutism
Call	the text is not an argument but an encouragement to act or think in a particular way	Slogans, Appeal to time, Conversation killer
Manipulative wording	specific language is used or a statement is made that is not an argument and which contains words/phrases that are either non-neutral, confusing, exaggerating, etc., in order to impact the reader, for instance emotionally	Loaded language, Repetition, Exaggeration or minimisation, Obfuscation - vagueness or confusion
Attack on reputation	an argument whose object is not the topic of the conversation, but the personality of a participant, his experience and deeds, typically in order to question and/or undermine his credibility	Name calling or labeling, Doubt, Guilt by association, Appeal to Hypocrisy, Questioning the reputation

Table 2.4: Persuasion technique taxonomy from Piskorski et al. (2023b). The six coarse-grained techniques are subdivided into 23 fine-grained ones.

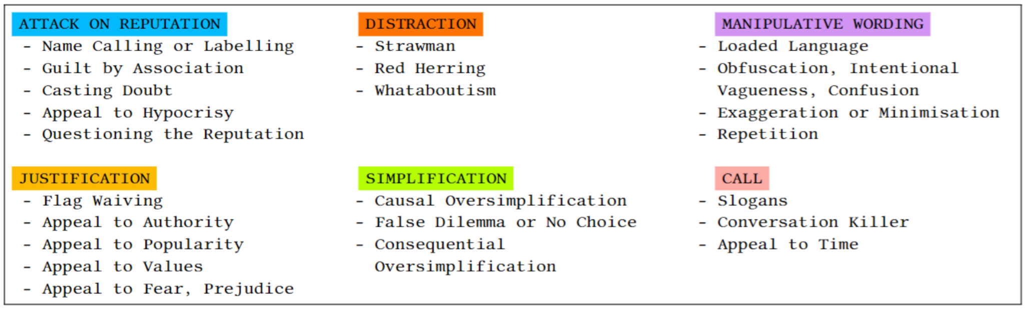
Chapter 3

Problem Definition

This chapter outlines the core problem addressed by this thesis: the automatic detection of persuasion techniques in the news in a multilingual context, a critical challenge in combating the spread of propaganda in digital media. With the increasing sophistication of computational propaganda, identifying the subtle mechanisms used to influence public opinion has become a pressing issue. The relevance of participating in shared tasks such as CheckThat! Lab 2024 lies in their structured and collaborative approach, which provides an invaluable platform for benchmarking models, addressing multilingual complexities, and uncovering gaps in current methodologies. By framing the problem within the context of a competitive and gamified shared task, this chapter sets the stage for a detailed discussion of the task, dataset, annotation schema, and evaluation criteria, highlighting the intricacies and challenges of detecting persuasion techniques in a multilingual and dynamic media landscape.

3.1 Overview of the Shared Task

The *CheckThat! Lab 2024 Task 3 on Persuasion Techniques* at CLEF 2024 is the latest collective effort to advance the state of the art in automatic persuasion technique detection (Piskorski et al., 2024). Participants were given a set of news articles in multiple languages and the two-tier persuasion technique taxonomy introduced in SemEval-2023 (Piskorski et al., 2023b) (Table∼2.4). At the top level, there are 6 coarse-grained types of persuasion techniques: *Attack on reputation, Justification, Simplification, Distraction,*



ATTACK ON REPUTATION
- Name Calling or Labelling
- Guilt by Association
- Casting Doubt
- Appeal to Hypocrisy
- Questioning the Reputation

DISTRACTION
- Strawman
- Red Herring
- Whataboutism

MANIPULATIVE WORDING
- Loaded Language
- Obfuscation, Intentional Vagueness, Confusion
- Exaggeration or Minimisation
- Repetition

JUSTIFICATION
- Flag Waiving
- Appeal to Authority
- Appeal to Popularity
- Appeal to Values
- Appeal to Fear, Prejudice

SIMPLIFICATION
- Causal Oversimplification
- False Dilemma or No Choice
- Consequential Oversimplification

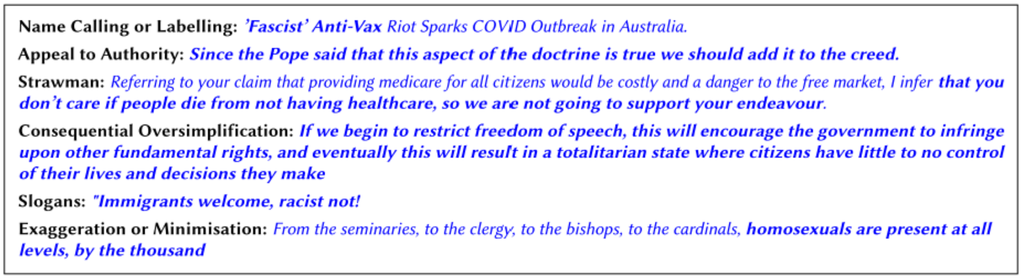
CALL
- Slogans
- Conversation Killer
- Appeal to Time

Figure 3.1: Persuasion technique taxonomy of CheckThat! Lab 2024 as presented to the participants. The taxonomy is the same of SemEval-2023 Task 3.

Call, and *Manipulative wording*. These six types are further subdivided into 23 fine-grained techniques, including logical fallacies (e.g., straw man, red herring, bandwagon) and emotional manipulation techniques (e.g., loaded language, appeal to fear, name calling) that might be used to support flawed argumentation (Figure~3.1). Figure~3.2 provides one example of persuasion technique per main category.

The goal of this task was to locate and label specific segments (spans) of text where each technique occurs. A unique aspect of this problem is that these labeled spans can overlap, meaning that a single portion of text might belong to more than one category simultaneously. To handle this, the problem is framed as a multi-label sequence-tagging task. In this setup, each word or token in the text can be assigned multiple labels to reflect the overlapping categories. The evaluation metric used was the micro-averaged F_1 score, a measure of precision and recall calculated across all labels, modified to account for partial matching between the spans. This ensures a fairer assessment of the model's performance when the predicted spans do not exactly align with the true spans.

The task covered highly-debated topics in the media, e.g., the Isreali–Palestinian conflict, the Russia–Ukraine war, climate change, COVID-19, and abortion. For training and development the participants were provided with the same dataset of the SemEval-2023 task 3 (Piskorski et al., 2023b), which covers nine languages: English, German, Georgian, Greek, French, Italian, Polish, Russian, and Spanish. For testing, the participants were provided with an entirely new dataset that covers five languages: Arabic, Bulgarian, English,



Name Calling or Labelling: ***’Fascist’ Anti-Vax*** *Riot Sparks COVID Outbreak in Australia.*

Appeal to Authority: ***Since the Pope said that this aspect of the doctrine is true we should add it to the creed.***

Strawman: *Referring to your claim that providing medicare for all citizens would be costly and a danger to the free market, I infer* ***that you don’t care if people die from not having healthcare, so we are not going to support your endeavour****.*

Consequential Oversimplification: ***If we begin to restrict freedom of speech, this will encourage the government to infringe upon other fundamental rights, and eventually this will result in a totalitarian state where citizens have little to no control of their lives and decisions they make***

Slogans: ***"Immigrants welcome, racist not!***

Exaggeration or Minimisation: *From the seminaries, to the clergy, to the bishops, to the cardinals,* ***homosexuals are present at all levels, by the thousand***

Figure 3.2: Examples of text snippets with persuasion techniques. The text fragments highlighted in bold are the actual text spans annotated.

Portuguese, and Slovene. English is the only language for which both training/development and test data existed. For Arabic, the dataset covers forteen broad topics such as news, politics, health, social, sports, arts and culture, religion, science and technology, human rights, and lifestyle.

3.2 Data

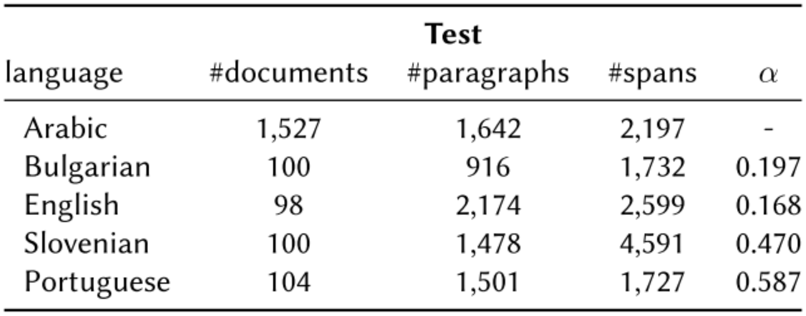
Annotation Process Each language was annotated by a team of annotators selected by the shared task’s organizers. Each annotator was required to be fluent in the language used to perform such annotations, and the language leaders met regularly to discuss difficult cases with more experienced annotators. For all languages but Arabic, each document was annotated by two annotators, and one curator reconciled the annotations. For the Arabic test dataset, each paragraph was annotated by three annotators, and two curators consolidated the annotations. The annotators were first given the comprehensive annotation guidelines (Piskorski et al., 2023a), then were further trained using two sets of flashcards of increasing complexity, and lastly had to annotate and discuss with expert annotators five test documents whose ground-truth annotations were known.
The overall Inter-Annotator Agreement (IAA) reported by the organizers as measured by the Krippendorff’s α is of 0.404, which is lower than the recommended value of 0.667 (Krippendorff, 2004). However, the organizers point out that one has to take into account that this measures coherence before curation, and that significant steps have been taken in order to improve the quality of the curated data: they clustered the annotations based

	Training		Development	
language	#documents	#spans	#documents	#spans
English	536	9,002	54	1,775
French	211	6,831	50	1,681
German	177	5,737	50	1,904
Italian	303	7,961	61	2,351
Polish	194	3,824	47	1,491
Russian	191	4,138	72	944
Georgian	-	-	29	218
Greek	-	-	64	691
Spanish	-	-	30	546

Table 3.1: Training and development CheckThat! Lab 2024 data statistics.

on their semantic similarity, which allowed them to flag outliers for review and to spot and discuss cross-lingual disagreements. Such disagreements were either due to individual annotator differences or to a more fundamental different understanding of techniques' definitions across language-specific annotation teams. Furthermore, the organizers compared the distribution of labels to spot obvious cross-lingual inconsistencies, alphabetically sorted texts in order to make it easy to spot similar texts with different labels, and the most experienced annotators did random checks. After adopting these quality and coherence assurance measures, when ignoring the Loaded Language and Name-Calling/Labelling classes (the most frequent), the α value went from 0.279 to 0.284, and when considering them it increased from 0.608 to 0.611.

Statistics The overall statistics about the training and the development datasets are provided in Table∼3.1. Regarding the test data, for this task the organizer created new labeled datasets for Arabic, Bulgarian, English, Portuguese, and Slovene. With the exception of the latter, this shared task was the first application of the framework for annotating persuasion techniques for the mentioned languages. News selection was delegated to the teams responsible for their respective languages, but they were expected to include a variety of topics, news genres, and political stances, in addition to selecting texts where a high prevalence of persuasion techniques was to



		Test		
language	#documents	#paragraphs	#spans	α
Arabic	1,527	1,642	2,197	-
Bulgarian	100	916	1,732	0.197
English	98	2,174	2,599	0.168
Slovenian	100	1,478	4,591	0.470
Portuguese	104	1,501	1,727	0.587

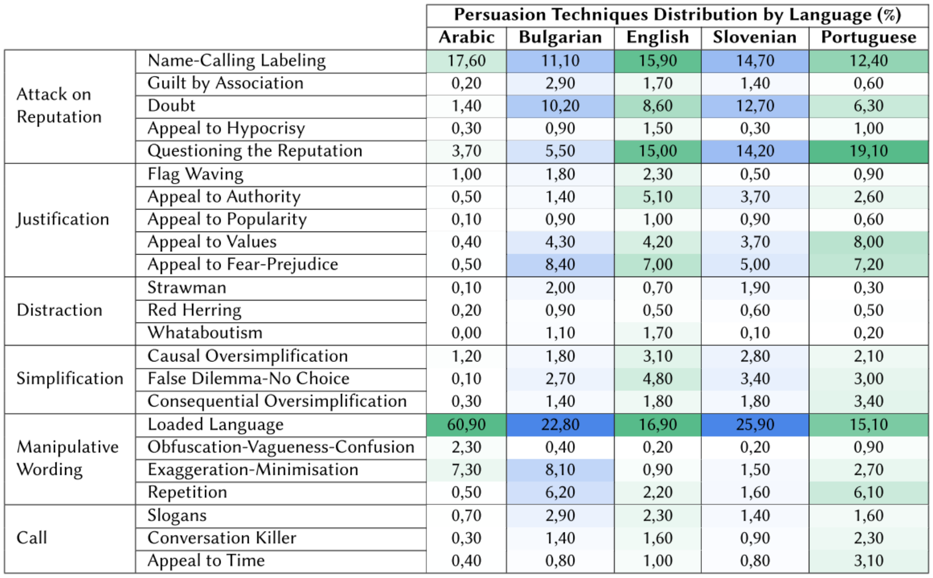
Table 3.2: Test CheckThat! Lab 2024 data statistics.

be expected. To allow for comparability with previous datasets, the topics of the Russia–Ukraine war, climate change, COVID-19, and abortion were covered in all test datasets except for Arabic. In addition, a new topic, the Israeli–Palestinian conflict, was added. Statistics about the test dataset can be found in Table∼3.2.

Overall, the most commonly annotated persuasion technique in the test dataset was *Loaded Language*, followed by *Name-Calling/Labelling, Casting Doubt* and *Questioning the Reputation*, although the specific distribution varies across the datasets. The share of annotated persuasion technique classes across the test datasets is presented in Table∼3.3. The distribution of the frequency of the persuasion techniques in the test dataset is to some degree similar and comparable to the datasets used in the SemEval-2023 Task on persuasion techniques, i.e. *Loaded Language* and *Name-Calling/Labelling* are the two most prevalent fine-grained techniques, whereas *Manipulative Wording* and *Attack on Reputation* are the two coarse-grained persuasion technique categories with the highest share in both datasets.

3.3 Evaluation

The task is defined as a multi-label multi-class sequence tagging problem. In such tasks, traditional evaluation metrics (e.g. Exact Match and F_1) tend to be too strict when scoring since they are based on exact matching. During the annotation process, the organizers noted that most of the time there was agreement between annotators on the technique, but the spans differed slightly. Assuming that from the end-user perspective partial matches with



		Persuasion Techniques Distribution by Language (%)				
		Arabic	Bulgarian	English	Slovenian	Portuguese
Attack on Reputation	Name-Calling Labeling	17,60	11,10	15,90	14,70	12,40
	Guilt by Association	0,20	2,90	1,70	1,40	0,60
	Doubt	1,40	10,20	8,60	12,70	6,30
	Appeal to Hypocrisy	0,30	0,90	1,50	0,30	1,00
	Questioning the Reputation	3,70	5,50	15,00	14,20	19,10
Justification	Flag Waving	1,00	1,80	2,30	0,50	0,90
	Appeal to Authority	0,50	1,40	5,10	3,70	2,60
	Appeal to Popularity	0,10	0,90	1,00	0,90	0,60
	Appeal to Values	0,40	4,30	4,20	3,70	8,00
	Appeal to Fear-Prejudice	0,50	8,40	7,00	5,00	7,20
Distraction	Strawman	0,10	2,00	0,70	1,90	0,30
	Red Herring	0,20	0,90	0,50	0,60	0,50
	Whataboutism	0,00	1,10	1,70	0,10	0,20
Simplification	Causal Oversimplification	1,20	1,80	3,10	2,80	2,10
	False Dilemma-No Choice	0,10	2,70	4,80	3,40	3,00
	Consequential Oversimplification	0,30	1,40	1,80	1,80	3,40
Manipulative Wording	Loaded Language	60,90	22,80	16,90	25,90	15,10
	Obfuscation-Vagueness-Confusion	2,30	0,40	0,20	0,20	0,90
	Exaggeration-Minimisation	7,30	8,10	0,90	1,50	2,70
	Repetition	0,50	6,20	2,20	1,60	6,10
Call	Slogans	0,70	2,90	2,30	1,40	1,60
	Conversation Killer	0,30	1,40	1,60	0,90	2,30
	Appeal to Time	0,40	0,80	1,00	0,80	3,10

Table 3.3: Distribution of persuasion technique labels in test dataset by language (percent). Color intensity represents the relative frequency of labels for each language.

significant overlap could be considered as equally good as exact matches, the organizers proposed an adjustment to the F$_1$-score to account for partial span matching and address the limitations of exact matching scorers. To understand the adjustment, let

- $P = \{p_1, \ldots, p_n\}$ be the set of predictions for one article, $p \in P$ is a generic prediction which is represented as an ordered triple $\langle span_{start}, span_{end}, label \rangle$

- $G = \{g_1, \ldots, g_m\}$ be the set of gold labels for one article, $g \in G$ is a generic gold label which is represented as an ordered triple $\langle span_{start}, span_{end}, label \rangle$

- $L : (p, g) \longrightarrow \{0, 1\}$ is a function that measures the similarity of the labels of p and g:

$$L(p, g) = \begin{cases} 1, & \text{if the labels of } p \text{ and } g \text{ are identical} \\ 0, & \text{otherwise} \end{cases}$$

- $I : (p, g) \longrightarrow [0, 1]$ is a function that measures the overlap rate of the spans of p and g:

$$I(p,g) = \begin{cases} 1, & \text{if } \frac{|p \cap g|}{|g|} \geq 0.5 \text{ and } |p| \leq 2 \cdot |g| \\ \frac{|p \cap g|}{|g|} \in (0,1), & \text{if } \frac{|p \cap g|}{|g|} \in (0, 0.5) \text{ and } |p| \leq 2 \cdot |g| \\ \frac{|p \cap g|}{|p|} \in (0,1), & \text{if } \frac{|p \cap g|}{|g|} \in (0, 1) \text{ and } |p| > 2 \cdot |g| \text{ and } |p| \leq 4 \cdot |g| \\ 0, & \text{otherwise} \end{cases}$$

- $S : (p, g) \longrightarrow [0, 1]$ is a similarity function between spans p and g. It is calculated as:

$$S(p, g) = L(p, g) \cdot I(p, g)$$

Each possible case is mapped into True Positive (T_p), False Positive (F_p), and False Negative (F_n) values, and then the standard, micro- and macro-averaged F_1 score is computed for the sample and for each persuasion technique.

Baseline System The baseline provided by the task organizers was an XLM-RoBERTa-base token classification model (Conneau et al., 2020) in a zero-shot setting. The baseline followed a simple, heuristic approach: for each token, the model predicted the classes with a given probability threshold, and then adjacent tokens with the same class were merged in a single span.

Chapter 4

Experiments

This chapter details the experiments conducted to develop and evaluate a system for detecting persuasion techniques in the news in a multilingual context, as part of the participation in the CheckThat! Lab 2024 Task 3 on Persuasion Techniques at CLEF 2024. The experiments aim to address the challenge of building a multilingual NLP system capable of identifying various persuasion techniques in news articles better than the state of the art. The chapter outlines two primary experimental efforts: the initial system developed for the shared task competition and a subsequent post-competition enhancement phase to refine its performance. These experiments involve preprocessing and augmenting the multilingual dataset and training advanced models for sentence- and token-level classification. Results from these experiments provide insights into the system's strengths and weaknesses, forming the basis for discussions on challenges and limitations.

4.1 Shared Task Experiment

4.1.1 Methodology

The pipeline proposed by our team for the system submitted to the *CheckThat! Lab 2024 Task 3 on Persuasion Techniques* at CLEF 2024 encompasses two modules: i) a data preprocessing and augmentation module and ii) a persuasion technique classification module. This subsection includes an overview of our system, while the following subsections contain further details.

In the first module, first we split the documents in the training dataset into

Language	**docs**	**sentences**	**Language**	**docs**	**sentences**
English	536	24,514	Italian	303	7,917
French	211	6,298	Polish	194	9,612
German	177	5,667	Russian	191	5,900
			Total	**1,612**	**59,908**

Table 4.1: Statistics of the training data in terms of number of documents and sentences after preprocessing and before augmentation.

sentences to facilitate further processing. Then, for the data augmentation, we first translate the English training sentences into the other training languages via machine translation (MT), and then use a BERT-based model fine-tuned on a word-alignment task to project the gold labels from the source to the target sentences.

The second module of the pipeline refers to the persuasion technique classification: two separate BERT-based models, henceforth referred to as sequence classifier and token classifier. The binary sequence classifier is trained to classify individual sentences as containing a persuasion technique or not. The token classifier is a series of 23 token-level classifiers, one per persuasion technique. Leveraging our multilingual MT data augmentation strategy, we trained a set of multilingual models and used them to infer on all languages of a holdout validation set for internal experimentation and of the official test set from the competition. Thus, our system aims at pointing out the exact occurrences of specific persuasion techniques in the news. For reproducibility purposes, we release our code and pre-processed data as described above[1].

4.1.2 Data Preprocessing and Augmentation

Regarding the first module of the pipeline, we first observe that the documents are too long to feed in input to the hereby-used models (BERT-based models that can only handle 512 tokens in input, details in Section∼4.1.3). Therefore, we generate smaller training samples by splitting documents at the sentence level, obtaining 59,908 total gold training sentences, as indicated in Table∼4.1. Prior to training, we split the obtained sentence dataset 80/20 into training and validation instances.

[1] https://github.com/giorluca/checkthat24_DIT

The next step after preprocessing is data augmentation. In order to increase the amount of available training data, we augment the training sentences via MT and label projection. MT is carried out by translating the dataset sentence by sentence from English to the other training languages with the NLLB 3.3B model (Costa-jussà et al., 2022). Following Nagata et al. (2020), the gold labels for the persuasion techniques are then projected onto the translated text by using mDeBERTa models (He et al., 2021) trained on a word-alignment task with a question-answering classifier head. The task of label projection is defined as follows: given a source sentence A with characters $a_i \in A$, and its translated target sentence B with characters $b_j \in B$, and an alignment between spans $a_{i,i+k}$, labeled as C, and $b_{j,j+l}$, with $i < j \in \mathbb{N}$, $k, l > 0 \in \mathbb{N}$, assign the label C to the span $b_{j,j+l}$ (Jain et al., 2019). In other words, given a source span, the model is tasked to find the equivalent span in the translated text. In order to train these word-alignment models, we use XL-WA (Martelli et al., 2023), a multilingual word-alignment dataset built from WikiMatrix (Schwenk et al., 2021). The dataset has a balanced domain distribution and features 14 EN-XX language combinations. Its training set is composed of silver labels automatically generated, while the development and test sets are manually annotated. We align each source-target combination of machine-translated data (EN-IT, EN-ES, EN-RU, EN-SL, EN-BG, EN-PT), where English is always the source gold data, with a different word-alignment model, trained on the specific language combination contained in XL-WA.

More specifically, in the approach proposed by Nagata et al. (2020), the source word to be aligned is enclosed within rarely used characters, such as '•', and the model is fed both the source sequence A and the target sequence B simultaneously. The input to the model at the token level is structured as follows:

$$\begin{aligned}&\texttt{[CLS]}\,\alpha_1, \ldots, tok(\bullet), \alpha_i, \ldots, \alpha_{i+k}, tok(\bullet), \ldots, \alpha_m\\&\texttt{[SEP]}\,\beta_1, \ldots, \beta_j, \ldots, \beta_{j+l}, \ldots, \beta_n\ \ \texttt{[SEP]}\end{aligned}$$

Here, the source word to be aligned is represented by the tokens $\alpha_i, \ldots, \alpha_{i+k}$, where $\alpha_i \in tok(A)$. The model is then tasked with predicting the tuple (β_j, β_{j+l}), where $\beta_j \in tok(B)$, which denotes the boundary indices of the aligned word in the target sequence.

For each language combination involved in the data augmentation process, we train our models for up to 3 epochs on each of XL-WA's languages with a batch size of 16. The optimizer's learning rate is set to 3×10^{-4}, and ϵ is 10^{-8}. We select the best model based on the Exact metric E (Rajpurkar et al., 2018):

$$E = \frac{\sum_i^n exact(p_i, g_i)}{\|preds\|} \ , \tag{4.1}$$

where $preds$ is a list of predictions and $exact(p_i, g_i)$ is the Kronecker delta:

$$exact(p_i, g_i) = \begin{cases} 1, & \text{if } p_i = g_i, \\ 0, & \text{if } p_i \neq g_i. \end{cases} \tag{4.2}$$

Before computing E, we lowercase and strip the predicted and gold strings p_i and g_i of excess punctuation and spacing.

Doing this, we obtain synthetic annotated data in the target languages. Ultimately, departing from the original 24,514 gold English sentences indicated in Table~4.1, we generate the same amount for each of the six target languages, for a total of 147,084 extra training sentences. Thus, the total number of training instances amounts to 206,992.

Prior to training for token classification, we preprocess and label the data using the BIO annotation scheme (Ramshaw and Marcus, 1999). In this scheme, the first word of an entity is assigned a `B-{class}` (beginning) label, subsequent words are assigned an `I-{class}` (inside) label, and words not part of any entity are assigned an `O` (outside) label. We follow established methodology by ignoring subword tokens when calculating cross-entropy loss[2].

4.1.3 Persuasion Technique Classification

Sequence classifier Upon training, we feed the sequence classifier a balanced subsample of the sentence dataset, obtained as per Section~4.1.2. Specifically, we take all sentences containing at least one persuasion technique (considered positive instances) and sample an equal number of negative instances (which contain no persuasion technique) from the rest of the training set.

[2] `https://huggingface.co/docs/transformers/en/tasks/token_classification`

Token classifiers Since the 23 token classification models are tailored specifically to each PT, we train them on sentences where only one PT is kept at a time. This means that if a sentence contains a persuasion technique which the model is not supposed to learn to predict, we set the tokens relative to that persuasion technique to the outside `O` label. Just like for the sequence classifier, we balance positive and negative instances for training.

For both the sequence classifier and the token classifier, we set the optimizer's learning rate at 5×10^{-5}, while ϵ is 10^{-8}. We train all models for up to 10 epochs with a patience of 2 epochs, keeping the model with the highest performance on the validation set.

Reducing False Positives As we are using 23 separate token classifiers, we observe that the number of predictions being produced ends up being very large. Since the submission website for the task only accepts TXT files of up to 800 KBytes and our token classifiers produce too many predictions, our full outputs are not suitable for submission in most languages. As such, we opt for reducing the number of positive predictions in order to adhere to the submission size limit.

To accomplish this, during training we use a modified, weighted version of the cross-entropy loss function. Specifically, we empirically assign a weight of 0.5 to the `O` majority class (label 0) and a weight of 2.0 to the minority `B` and `I` classes (labels 1 and 2). This weighting ensures that the model pays more attention to correctly predicting the minority classes, thus reducing the overall number of positive predictions.

When computing the evaluation metrics, we also apply a threshold to the model's predictions. We use the softmax function to convert the logits into probabilities. Then, we set any probability below the threshold of 0.9 to zero before determining the predicted labels.[3] This means that the model only makes a prediction if it is at least 90% confident, reducing the number of false positives. We did not experiment with any other parameters, besides function loss weights and the prediction threshold. Finally, since six of the token classifiers (i.e. those trained to identify the techniques *Appeal to Values, Red Herring, Appeal to Popularity, Obfuscation-Vagueness-Confusion, Straw Man* and *Whataboutism*) obtained an F_1 score of 0 on their class subset of the validation partition obtained from splitting the training set, we

[3] We attempted different thresholds by increments of 0.1 until the submission files were small enough for submission.

exclude the predictions produced by those models.

Inference During inference, we produce the submission predictions following a series of steps. First, after the models have produced their predictions, we set the token classifier's predictions to 0-tensors for those indices where the sequence classifier's predictions are 0. Then, we binarize the predictions to $\{0, 1\}$, with the original $\{1, 2\}$ labels mapping onto 1, and 0 mapping onto 0. Lastly, we assign a character span to each consecutive series of positive (1) predictions in the prediction tensor, based on the characters corresponding to each token.

4.1.4 Results and Discussion

The binary sequence classifier performs decently, with a macro F_1 of 0.757 on the holdout validation set which contains all training languages, obtained by splitting the training set 80/20 into training and validation data (Table∼4.2). Furthermore, as shown by the scores obtained in different experimental settings reported in Table∼4.3, the data augmentation more than doubles the token classifier's performance (from 0.168 to 0.336). On the other hand, class weighting and setting a decision threshold as high as 0.9, although necessary as shown above, lowers the best performance from 0.336 to 0.192. Since these preliminary tests conducted on the holdout validation split show that data augmentation improves the performance of the token classifier (from 0.168 to 0.336), even when class weighting and a decision threshold of 0.9 are set (from 0.168 to 0.192), we choose to adopt data augmentation also for the final system used to predict on the test set for submission. The rationale behind this decision is based on the assumption that a higher performance on the holdout validation set would also translate onto the test set.[4]
The results for our official test runs achieved by our whole system are shown in Table∼4.4. Our system performs better than the baseline across all languages, ranks first among the other participants for all except for Arabic, and performed better than the state of the art proposed by the organizers for two test languages out of five. For Arabic, Team Mela used a multilingual BERT

[4] Note that our official submission (last row in Table∼4.3) is not the best because it is constrained by the maximum size accepted by the submission website for the produced prediction file. Indeed, the used class weights and prediction threshold are applied in order to reduce the amount of predictions produced by the model.

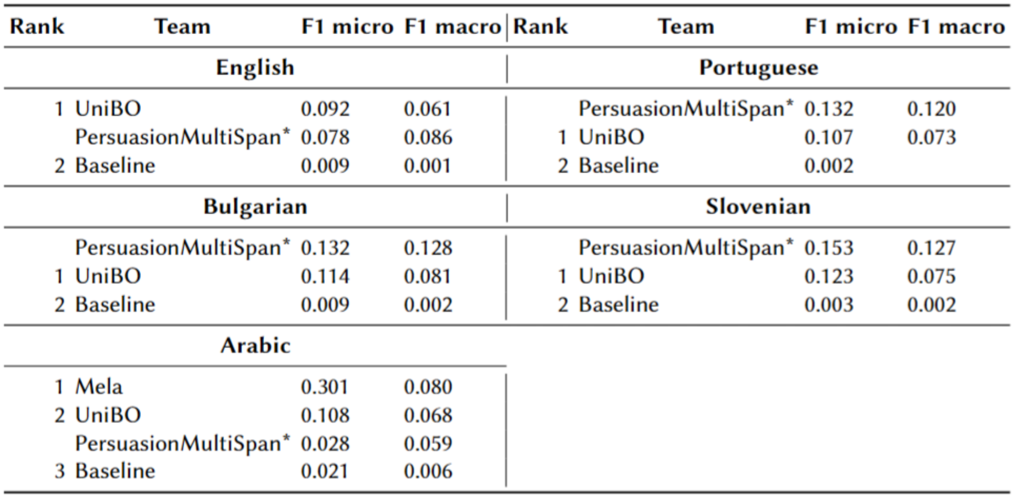
Class	P	R	F_1
0	0.802	0.662	0.726
1	0.734	0.850	0.788
Macro AVG	0.768	0.756	0.757

Table 4.2: Results obtained by the binary sequence mDeBERTa classifier on the holdout validation set obtained from the 80/20 split of the training set. The reported scores are achieved on all languages of the holdout validation set at once.

Data Aug.	Class Weighting	Threshold	Macro F_1
			0.168
✓			0.336
✓	✓	0.9	0.192

Table 4.3: Average results obtained by the 23 mDeBERTa token classifiers on the holdout validation set from the training set (80/20 split). The macro F_1 scores are achieved on all languages of the holdout validation set at once.

model which was pre-trained on data in both English and Arabic (Nabhani and Riyadh, 2024). Our system is competitive in all language settings, with micro average F_1 scores ranging from 0.092 for English to 0.123 for Slovenian, possibly showing hints of cross-lingual transfer ability when training the model on multi-lingual data and testing it on unseen languages.
For the sake of comparison to state-of-the-art solutions, the organizers developed and submitted (after the competition) a multi-lingual token-level multi-label classifier of persuasion techniques (referred to in Table∼4.4 with evaluation results with **PersuasionMultiSpan***) based on XML-RoBERTa (Conneau et al., 2020) trained on the SemEval-2023 corpus (Piskorski et al., 2023c), capable of processing arbitrarily long text using sliding window chunking with 50% overlap. This classifier achieves state-of-the-art results on the SemEval-2023 Task 3 test dataset (Piskorski et al., 2023b) for all six languages (oscillates around 1-3 rank across languages), both in terms of micro and macro F_1 scores. Further detail about this classifier can be found in Nikolaidis et al. (2024). It is worth noting that our system scores higher than this model in English and Arabic.



Rank	Team	F1 micro	F1 macro	Rank	Team	F1 micro	F1 macro
	English				**Portuguese**		
1	UniBO	0.092	0.061		PersuasionMultiSpan*	0.132	0.120
	PersuasionMultiSpan*	0.078	0.086	1	UniBO	0.107	0.073
2	Baseline	0.009	0.001	2	Baseline	0.002	
	Bulgarian				**Slovenian**		
	PersuasionMultiSpan*	0.132	0.128		PersuasionMultiSpan*	0.153	0.127
1	UniBO	0.114	0.081	1	UniBO	0.123	0.075
2	Baseline	0.009	0.002	2	Baseline	0.003	0.002
	Arabic						
1	Mela	0.301	0.080				
2	UniBO	0.108	0.068				
	PersuasionMultiSpan*	0.028	0.059				
3	Baseline	0.021	0.006				

Table 4.4: Official leaderboard results for all five languages, both in terms of micro and macro F_1. We include the two official participant runs (UniBO and Mela) and the baseline. The team marked with * is a post competition experiment from the organizers.

4.2 Post-competition Experiment

4.2.1 Methodology

The objective of the post-competition experiments was to enhance the performance of the system proposed at the CheckThat! Lab 2024. Most of what was discussed in Section~4.1 applies here, though several measures were taken to try and enhance the performance of the system. This subsection contains an overview of the approach, while in the following subsections more details are reported.
The main steps taken to try and enhance the performance of the system were:

1. Inject the machine-translated augmented data (as per Section~4.1.2) not only for the 23 token classifiers, but also for the binary sequence classifier

2. Further expand the training data for the system by integrating other suitable datasets from other initiatives

3. Experiment with hyperparameter optimization

Enhancing the Sequence Classifier To enhance the performance of the binary sequence classifier, several experiments in different settings and combinations thereof were conducted.

First, the whole machine-translated augmented data was provided to the model for training together with the shared task training set without changing anything else in comparison to the first experimental setting. Then, several learning rates were tested to further tune the performance. Furthermore, two additional datasets were integrated in the augmented training data, i.e. the news paragraph subset of the dataset from ArAiEval-2024 (Hasanain et al., 2024b) and the dataset from SemEval-2024 (Dimitrov et al., 2024), described in Section∼2.4. Finally, an ablation study was conducted to better understand to what extent each variable in the different experimental settings contributed to the best performance obtained. The settings in the ablation study included different combinations of training data and learning rates, such as i) no data augmentation + lower learning rate, ii) MT augmented data + higher/lower learning rates, iii) only the Italian and Russian sentences from the MT augmented data since they were the only languages from the augmented data that are also found in the 20% holdout validation set + higher/lower learning rates, and iv) MT augmented data + ArAiEval-2024 dataset (+ SemEval-2024 dataset).

Enhancing the Token Classifiers To try and enhance the performance of the 23 token classifiers, similar steps to those described in the previous paragraph for the sequence classifier were taken, except for the MT augmented data injection, which had already been done in the context of the shared task.

The experiments to enhance the token classifiers are rather straightforward, and included further augmenting the training data by incorporating the news paragraph subset of the dataset from ArAiEval 2024 (Hasanain et al., 2024b) and the dataset from SemEval 2024 (Dimitrov et al., 2024).

4.2.2 Results and Discussion

The results presented in this section are those obtained by testing the fine-tuned models on the same 20% holdout validation set as described in Section∼4.1.2.

Unfortunately, despite trying to get in contact with the shared task organizers, the public post-competition leaderboard, which for shared tasks is sometimes available after the event takes place to allow for the development and testing of better models even after the official runs, was closed and new submissions were not (yet) possible at the time of writing. Therefore, the results reported in this section are to be understood as preliminary results obtained in the context of internal experimentation, and cannot and should not be compared to the results obtained in the official shared task runs found in Table~4.4, but rather to those reported in Table~4.2 for the sequence classifier and Table~4.3 for the token classifier. However, with the aim of avoiding having models obtain an F_1 score of 0 (as described in Section~4.1.3), a training code and validation logic optimization and revision for the token classifier was conducted, which preceded any other attempts at enhancing the performance. This revision led to discovering some of the errors in the initial code that hindered some models and made them fail at predicting their class, and correcting these errors led to an increase in the average performance. Therefore, when interpreting the results of the post-competition experiment, the performance in internal experimentation of the shared task experiment (Table~4.3) is ignored, and the attempts at enhancing the models will be compared against the performance after the code optimization and revision, which here is considered as a substitute of the performance of the first experiment, rather than an attempt at enhancing the model in the post-competition experiment (Table~4.6).

The results for the enhanced binary sequence classifier can be found in Table~4.5. The first noteworthy phenomenon is that any setting in which any data augmentation was conducted without lowering the learning rate from the value of the baseline (5e-5 as in the first experiment, #3 in Table~4.5) led to a drastic drop in performance, with macro averaged F_1 scores as low as 0.334, as can be seen in settings #1 and #2. This is not easily explainable and further testing is needed to assess the reasons behind this unexpected behavior. However, a great increase in performance was observed with the injection of the full MT augmented data paired with a learning rate of 2e-5 (#8). The macro averaged F_1 score increased from 0.757 of the baseline to 0.847. At this point, an ablation study was conducted to better understand to what extent each variable in the different experimental settings contributed to this performance. As shown in #4, lowering the learning rate from 5e-5 to 2e-5 alone already increased the performance from 0.757 to 0.771. When keeping this learning rate, injecting MT augmented Italian and Russian data

#	Data aug.	LR	Macro F_1↓
1	MT	5e-5	0.334
2	MT (IT and RU only)	5e-5	0.335
3	-	5e-5	0.757
4	-	2e-5	0.771
5	MT (IT and RU only)	2e-5	0.784
6	MT + ArAiEval24-news	2e-5	0.833
7	MT + ArAiEval24-news + SemEval-24	2e-5	0.842
8	MT	2e-5	0.847

Table 4.5: Results obtained in the post-competition experiment by the binary mDeBERTa sequence classifier in different settings on the holdout validation set obtained from the 80/20 split of the training set. The reported scores are achieved on all languages of the holdout validation set at once. Sorted by macro averaged F_1 score. #3 comes from the shared task experiment, and represented the baseline for improvement.

(#5) further increased the performance to 0.784. The need to test data augmentation with only these languages was borne out of the observation that the performance obtained in #8 might have been only thanks to those languages, since they were the only two from the augmented data that are also found in the 20% holdout validation set. However, setting #8 allows us to reject this observation, showing that adding the MT augmented data in the other languages, i.e. Slovenian, Portuguese, Spanish and Bulgarian, helped increase the performance. It is interesting to observe that these 4 languages are not found in the 20% holdout validation set. Therefore, this might be a clear sign of proper cross-lingual transfer learning (Conneau et al., 2020) between related languages (Lin et al., 2019). To further confirm that this is the case, settings #6 and #7 must be considered. In these settings, there is little to no difference in performance compared to setting #8, despite augmenting the training data also with the news paragraph subset of the dataset from ArAiEval-2024 in setting #7 and additionally with the dataset from SemEval-2024 in setting #8. Keeping in mind that these results are based on the 20% holdout validation set, no cross-lingual transfer learning is observed in setting #7, most probably due to the fact that the new dataset contains exclusively samples in Arabic, a language not related to other ones in the training and validation sets. Regarding the dataset from SemEval-2024, it also did not contribute to the performance on the validation set,

#	Data Aug.	CW	TS	Macro F_1↓
1	-	-	-	0.266
2	MT + ArAiEval24-news	✓	0.9	0.327
3	MT	✓	0.9	0.331
4	MT	-	-	0.342
5	MT + ArAiEval24-news + SemEval-24	✓	0.9	0.349

Table 4.6: Average results in the post-competition experiment obtained by the 23 mDeBERTa token classifiers on the holdout validation set from the training set (80/20 split). The macro F_1 scores are achieved on all languages of the holdout validation set at once. Sorted by F_1 score. #1, #3 and #4 are the performances of the shared task experiment with the revised training code, with #3 representing the baseline for improvement.

most probably due to the difference in textual genres, since it contains internet memes and not news. However, it must be noted that the best model in a real-world scenario would most probably be setting #6 or #7, because they can also be used for inference on news in Arabic (and maybe related languages) and internet memes (albeit the performance is unknown due to the impossibility to make submissions on the post-competition shared task leaderboard, which computes the results on the test set, which contains the Arabic language).

The results for the enhanced multi-label token classifier can be found in Table∼4.6. #1, #3 and #4 are the performances of the shared task experiment with the revised training code, with #3 representing the baseline for improvement, since it is the model that is suitable for submission on the shared task leaderboard. It is important to emphasize how different these performances are compared to the ones before the code revision: at this point, some of the observations made in the first experiment, such as the impact of the data augmentation (i.e. from #1 to #4) and of the class weighting and threshold (i.e. from #4 to #3) on the performance of the models do not stand true anymore, or at least should be downsized. Indeed, also in this second experiment the data augmentation helps increase the performance, but by a smaller margin than the first experiment. Furthermore, the class weighting and threshold lead to a minimal performance decrease.

Regarding the attempts at enhancing the model, it is clear that further augmenting the training data with the two additional datasets did not lead to an increase of the performance on the 20% holdout validation set. However, the same observations made for the sequence classifier in the shared task experiment stand true, i.e. no cross-lingual transfer learning is observed in setting #2, most probably due to the fact that the new dataset contains exclusively samples in Arabic, a language not related to other ones in the training and validation sets, while regarding the dataset from SemEval 2024, it also did not contribute to the performance on the validation set (#5), most probably due to the difference in textual genres, since it contains internet memes and not news. However, as for the sequence classifier, it must be noted that the best model in a real-world scenario would most probably be setting #2 or #5, because they can also be used for inference on news in Arabic (and maybe related languages) and internet memes (however further testing is required).

Chapter 5

Conclusion

This thesis introduced the reader to the concept of propaganda and persuasion techniques, with a focus on computational propaganda in news and in multilingual contexts, providing an in-depth, thorough overview of the scientific literature published on this subject and highlighting gaps, challenges, limitations, and ethical considerations. The main aim of this work has been developing a system for the automatic detection of persuasion techniques in the news in a multilingual setting that can perform better than state-of-the-art approaches. Two versions of the proposed system were developed: the first version was developed in the context of the participation of the UniBO team (Gajo et al., 2024) to the *CheckThat! Lab 2024 Task 3 on Persuasion Techniques* (Piskorski et al., 2024) at CLEF 2024. The task consisted in detecting 23 persuasion techniques at the fragment-level in online media and covered highly-debated topics e.g., the Isreali–Palestian conflict, the Russia–Ukraine war, climate change, COVID-19, and abortion. The second version was developed months later, as an attempt to enhance the performance of the first version.

The first version of the system, which was submitted for the official run of the shared task, performed better then the baseline across all languages, ranked first among the other participants for all except for Arabic, and performed better than the state-of-the-art approach proposed by the organizers for two test languages out of five. The system is competitive in all language settings, with micro average F_1 scores ranging from 0.092 for English to 0.123 for Slovenian, possibly showing hints of cross-lingual transfer ability when training the model on multi-lingual data and testing it on unseen languages. The second version of the system performed slightly better than the first version in internal experimentation, although it was not possible to test it against the official test set and directly compare the two systems. However, the methods employed in the post-competition experiment and the considerations made in the results contribute to a potential better real-world performance in the future.

The results of the two experiments contribute to this field by highlighting the major challenges of the task and the limitations of contemporary approaches, but also what might be potential solutions or hints toward better performing systems.

Challenges and Limitations Automatic propaganda detection is still in its early stages and there are several challenges to overcome and gaps in the literature to be filled.

One of the main challenges in propaganda detection is modelling its multimodal nature. According to Da San Martino et al. (2021), “text is not the only way to convey propaganda. Sometimes, pictures convey stronger messages than texts, as for certain political memes. Thus, it is becoming increasingly necessary to analyze multiple modalities of data (e.g., images, videos, speech)”. Although some research has been done in this direction (e.g., Dimitrov et al. (2021a)), it is still one of the most promising research areas to investigate to fight online propaganda.

Another highly relevant challenge and research gap to be filled is the explainability of the models trained to detect propaganda and persuasion techniques. These models need to be able to provide motivations for the algorithmic decisions taken in order to be trusted by the end users (e.g., analysts carrying out media analysis, professionals in the journalistic practice or the general public) and to actually help them think more critically. However, most of the recent developments in propaganda and coordination detection, with notable exceptions such as Yu et al. (2021), are based on deep learning, which lacks explainability, and this remains an open problem.

An ever more relevant challenge to be addressed is the need to detect AI-generated propaganda. Recent advances in neural language models have made it difficult even for humans to detect synthetic text (Yang et al., 2018; Zellers et al., 2019). Da San Martino et al. (2021) predicted that automatically-generated propaganda would have surfaced in the near future, and recent studies confirm that propagandists today might (or indeed already do) benefit from the adoption of the same advanced, generative technolo-

gies that the NLP research community in turn employs to fight propaganda (Musser, 2023).
Multilinguality represents another relevant challenge. As shown in Section~2.2 and Section~2.4, several shared tasks have been organized and several datasets and models have been developed to address the issue of multilinguality in propaganda and persuasion technique detection. The most recent approaches employ models based on the transformer architecture that have been pretrained on multilingual data, covering hundreds of languages (such as XLM-RoBERTa or mDeBERTa, the base for our system). However, the performance of these models varies greatly between languages, especially comparing high-resource to low-resource ones.
One last challenge is the lack of very large, widely representative and balanced datasets with objective annotations. The annotations of the most recent datasets are created by humans to achieve a high enough level of quality, but this makes them intrinsically subjective, at least to some degree. Furthermore, the corpora always need to be updated with new domains and topics to ensure representativeness and adequate model performance on new texts, and consequently new expensive annotation efforts are needed. Moreover, in the available datasets it is often the case that some persuasion techniques are under-represented and some others are over-represented. Finally, Sprenkamp et al. (2023) point out that some techniques, such as *Bandwagon* or *Reductio ad hitlerum*, are likely not well-represented in some model's training data, which makes them harder to detect, even more so if they are under-represented in the dataset.

Ethical Considerations We believe that this research direction has a general positive broader impact, as it seeks to advance the state of the art for more effective detection of propaganda and persuasion techniques, which we consider to be beneficial for the society. However, we acknowledge its limitations and state here our ethical considerations, which overlap to a large extent with those of the authors of the works discussed in Chapter~2. First of all, misclassifications due to the low performance of our model (and, absolutely speaking, of all state-of-the-art models today) might have a negative impact on end users and on the sources of the news analyzed. Potential end users, such as professionals and practitioners in the journalistic world as well as laypeople, might be influenced to believe something that strays from the objective content of the news, users might over-rely on these systems, and faulty narratives might emerge on the basis of the alleged objectivity provided by our predictive models. Misclassifications might even harm the credibility of this research direction. Furthermore, misclassifications, while highlighting false positives, might obscure actual instances of persuasion techniques. Finally, misclassifications might also harm the sources of the news, which reputation is at stake. This also calls for discussion on responsibility and accountability on matters of artificial intelligence and algorithmic journalism.

As already stated above, the data used to train these models is intrinsically subjective and biased. This unintentional subjectivity can be traced mainly to the initial news article selection, to the topic representativeness, and to the human annotation. Therefore, it is important to keep in mind that also the decisions taken by the models trained on this data will be subjective and biased.

Finally, we acknowledge that disseminating models trained to detect persuasion techniques might lead to the adoption by malicious actors, which might employ these technologies to biasly and unfairly moderate news and other type of content based on biases that may or may not be related to demographics and other information within the text.

Future Work Mirroring the main limitations, some examples of potential avenues for future work are:

1. it is worth exploring ways to obtain high, consistent performance across a large range of high- and low-resource languages

2. more research is needed not only to better detect AI-generated text, but specifically AI-generated propaganda. For example, it would be interesting to see if and in which ways AI-generated and human generated propaganda differ

3. it might be worth it to investigate the existence of more explicit predictive features, including but not limited to textual and linguistic indicators or online dissemination patterns, that prove to be typical of propagandistic news, independently of news source and common across different languages, which could be leveraged alongside transformer models and data augmentation techniques

4. as shown in Section~2.2, some textual characteristics and dissemination patterns of propaganda online can be leveraged to automatically identify it. However, in addition to being able to classify individual documents as propaganda or single accounts as deceptive/coordinated, it would be useful to also provide information towards understanding the goals and the strategy of propaganda campaigns, since understanding these elements might help recognize the same patterns in the future in a different context and different propaganda campaigns. For example, identifying the propagandistic strategy of a certain group of malicious actors during a specific propaganda campaign might help to recognize the same strategy in a new, different campaign. Therefore, it is worth to explore ways to detect general propagandistic strategies of a group of actors rather than identifying propaganda when already disseminated (and, therefore, consumed)

Looking ahead, continued innovation in this domain will be essential for broadening the impact of automated persuasion detection, fostering greater media transparency, and empowering readers with tools for critical engagement. Automatic systems capable of confidently detect the presence of persuasion techniques in multilingual news might find practical application for journalists, media organizations and policymakers, for example in countering disinformation and propaganda campaigns. This work serves as a foundation for future studies aiming to deepen our understanding of persuasive communication in a globally interconnected media landscape.

Bibliography

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.551.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In Houda Bouamor, Hend Al-Khalifa, Kareem Darwish, Owen Rambow, Fethi Bougares, Ahmed Abdelali, Nadi Tomeh, Salam Khalifa, and Wajdi Zaghouani, editors, *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.wanlp-1.11.

Firoj Alam, Abul Hasnat, Fatema Ahmed, Md Arid Hasan, and Maram Hasanain. ArMeme: Propagandistic content in Arabic memes. *arXiv preprint arXiv:2406.03916*, 2024.

Zien Sheikh Ali, Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. AraFacts: the first large Arabic dataset of naturally occurring claims. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 231–236, 2021.

Christopher W Anderson. Towards a sociology of computational and algorithmic journalism. *New media & society*, 15(7):1005–1021, 2013.

Wissam Antoun, Fady Baly, and Hazem Hajj. AraBERT: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9, 2020.

Eugene Bagdasaryan and Vitaly Shmatikov. Spinning language models: Risks of propaganda-as-a-service and countermeasures. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 769–786. IEEE, 2022.

Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864, 2019. doi: https://doi.org/10.1016/j.ipm.2019.03.005.

Allan Bell. Language and the media. *Annual review of applied linguistics*, 15:23–41, 1995.

Gillian Bolsover and Philip Howard. Computational propaganda and political big data: Moving toward a more critical research agenda, 2017.

Alexandre Bovet and Hernán A Makse. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications*, 10(1): 7, 2019.

Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. Truth, lies, and automation. *Center for Security and Emerging technology*, 1(1): 2, 2021.

Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, 2015.

Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. Debot: Twitter bot detection via warped correlation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 817–822. IEEE Computer Society, 2016.

Aditya Chetan, Brihi Joshi, Hridoy Sankar Dutta, and Tanmoy Chakraborty. Corerank: Ranking to detect users involved in blackmarket-based collusive retweeting activities. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 330–338, 2019.

Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on dependable and secure computing*, 9(6):811–824, 2012.

Mark Coddington. Clarifying journalism's quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting. *Digital journalism*, 3(3):331–348, 2015.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747.

Henry T Conserva. *Propaganda techniques.* AuthorHouse, 2003.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*, pages 963–972, 2017.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav 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*, pages 162–170, 2019a.

Giovanni Da San Martino, Yu Seunghak, Alberto Barrón-Cedeño, Rostislav Petrov, Preslav Nakov, et al. 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)*, pages 5636–5646. Association for Computational Linguistics, 2019b.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, 2020a.

Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. Prta: A System to Support the Analysis of Propaganda Techniques in the News. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 287–293, 2020b.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4826–4832, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Barbara Diggs-Brown. *Strategic public relations: An audience-focused approach*. Cengage Learning, 2011.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. Detecting Propaganda Techniques in Memes. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.516.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, Giovanni Da San Martino, 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)*, pages 70–98, 2021b.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026, 2024.

Theresa Elstner, Frank Loebe, Yamen Ajjour, Christopher Akiki, Alexander Bondarenko, Maik Fröbe, Lukas Gienapp, Nikolay Kolyada, Janis Mohr, Stephan Sandfuchs, et al. Shared tasks as tutorials: A methodical approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15807–15815, 2023.

EUCommission. Action Plan against Disinformation. Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. *URL: https://ec. europa. eu/newsroom/dae/document. cfm*, 2022.

Guglielmo Faggioli, Nicola Ferro, Petra Galuščáková, and Alba García Seco de Herrera, editors. *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, CLEF 2024, Grenoble, France, 2024.

Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election. *Berkman Klein Center Research Publication*, 6, 2017.

Paolo Gajo, Luca Giordano, and Alberto Barrón-Cedeño. UniBO at CheckThat! 2024: Multi-lingual and Multi-label Persuasion Technique Detection in News with Data Augmentation and Sequence-Token Classifiers. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, pages 426–434, 2024.

Andreas Graefe. *Guide to automated journalism.* Tow Center for Digital Journalism Publications, Columbia University, 2016.

Christian Grimme, Mike Preuss, Lena Adam, and Heike Trautmann. Social bots: Human-like by means of human control? *Big data*, 5(4):279–293, 2017.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. Argotario: Computational Argumentation Meets Serious Games. In Lucia Specia, Matt Post, and Michael Paul, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-2002.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. Adapting serious game for fallacious argumentation to german: Pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Maram Hasanain, Fatema Ahmed, and Firoj Alam. Large language models for propaganda span annotation. *arXiv preprint arXiv:2311.09812*, 2023a.

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abed Freihat. ArAIEval Shared Task: Persuasion Techniques and Disinformation Detection in Arabic Text. In Hassan Sawaf, Samhaa El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha, Nizar Habash, Salam Khalifa, Amr Keleg, Hatem Haddad, Imed Zitouni, Khalil Mrini, and Rawan Almatham, editors, *Proceedings of ArabicNLP 2023*, pages 483–493, Singapore (Hybrid), December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.arabicnlp-1.44.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. Can GPT-4 Identify Propaganda? Annotation and Detection of Propaganda Spans in News Articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2724–2744, 2024a.

Maram Hasanain, Md. Arid Hasan, Fatema Ahmad, Reem Suwaileh, Md. Rafiul Biswas, Wajdi Zaghouani, and Firoj Alam. ArAIEval Shared Task: Propagandistic Techniques Detection in Unimodal and Multimodal Arabic Content. In Nizar Habash, Houda Bouamor, Ramy Eskander, Nadi

Tomeh, Ibrahim Abu Farha, Ahmed Abdelali, Samia Touileb, Injy Hamed, Yaser Onaizan, Bashar Alhafni, Wissam Antoun, Salam Khalifa, Hatem Haddad, Imed Zitouni, Badr AlKhamissi, Rawan Almatham, and Khalil Mrini, editors, *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 456–466, Bangkok, Thailand, August 2024b. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *The Eleventh International Conference on Learning Representations*, 2021.

Mick Healey. Linking research and teaching to benefit student learning. *Journal of Geography in Higher Education*, 29(2):183–201, 2005.

Renee Hobbs and Sandra McGee. Teaching about propaganda: An examination of the historical roots of media literacy. *Journal of Media Literacy Education*, 6(2):56–66, 2014.

S Hochreiter. Long Short-term Memory. *Neural Computation MIT-Press*, 1997.

Phillip N Howard and Bence Kollany. Bots, #strongerin and #brexit: Computational Propaganda during the UK-EU Referendum. *Social Science Research Network*, 2016.

Kristina Hristakieva, Stefano Cresci, Giovanni Da San Martino, Mauro Conti, and Preslav Nakov. The spread of propaganda by coordinated communities on social media. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 191–201, 2022.

Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton. Entity Projection via Machine Translation for Cross-Lingual NER. 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)*, pages 1083–1092, 2019.

Karen S Johnson-Cartee and Gary Copeland. *Strategic political communication: Rethinking social influence, persuasion, and propaganda.* Rowman & Littlefield, 2004.

Garth S Jowett and Victoria O'Donnell. *Propaganda & persuasion*. Sage publications, 2018.

Sarah Kreps, R Miles McCain, and Miles Brundage. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1):104–117, 2022.

Klaus Krippendorff. Content analysis: An introduction to its methodology (2nd thousand oaks, 2004.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Wei Li, Shiqian Li, Chenhao Liu, Longfei Lu, Ziyu Shi, and Shiping Wen. Span identification and technique classification of propaganda in news articles. *Complex & Intelligent Systems*, 8(5):3603–3612, 2022.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, 2019.

Shenghua Liu, Bryan Hooi, and Christos Faloutsos. Holoscope: Topology-and-spike aware fraud detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1539–1548, 2017.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019.

Abdurahman Maarouf, Dominik Bär, Dominique Geissler, and Stefan Feuerriegel. HQP: a human-annotated dataset for detecting online propaganda. *arXiv preprint arXiv:2304.14931*, 2023.

Federico Martelli, Andrei Stefan Bejgu, Cesare Campagnano, Jaka Čibe, Rute Costa, Apolonija Gantar, Jelena Kallas, Svetla Koeva, Kristina Koppel, Simon Krek, et al. XL-WA: a Gold Evaluation Benchmark for Word Alignment in 14 Language Pairs. 2023.

Michele Mazza, Stefano Cresci, Marco Avvenuti, Walter Quattrociocchi, and Maurizio Tesconi. RTbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM conference on web science*, pages 183–192, 2019.

Clyde Raymond Miller. *How to Detect Propaganda...: An Address Delivered at Town Hall, Monday, February 20, 1939*. Town Hall, Incorporated, 1939.

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*, 71:397–407, 2023.

Micah Musser. A cost analysis of generative language models and influence operations. *arXiv preprint arXiv:2308.03740*, 2023.

Sara Nabhani and Md Abdur Razzaq Riyadh. Mela at CheckThat! 2024: Transferring persuasion detection from English to Arabic - a multilingual BERT approach. In Faggioli et al. (2024).

Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. A Supervised Word Alignment Method based on Cross-Language Span Prediction using Multilingual BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 555–565, 2020.

Nikolaos Nikolaidis, Jakub Piskorski, and Nicolas Stefanovitch. Exploring the usability of persuasion techniques for downstream misinformation-related classification tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6992–7006, 2024.

Chad Nilep. “Code switching” in sociocultural linguistics. *Colorado research in linguistics*, 2006.

Leonardo Nizzoli, Serena Tardelli, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. Coordinated behavior on social media in 2019 UK general election. In *Proceedings of the international AAAI conference on web and social media*, volume 15, pages 443–454, 2021.

Andrew Patel and Jason Sattler. Creatively malicious prompt engineering. *WithSecure Intelligence*, 2023.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202.

Jakub Piskorski, Nicolas Stefanovitch, Valerie-Anne Bausier, Nicolo Faggiani, Jens Linge, Sopho Kharazi, Nikolaos Nikolaidis, Giulia Teodori, Bertrand De Longueville, Brian Doherty, et al. News categorization, framing and persuasion techniques: Annotation guidelines. *European Commission, Ispra, JRC132862*, 2023a.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. SemEval-2023 Task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, 2023b.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, 2023c.

Jakub Piskorski, Nicolas Stefanovitch, Firoj Alam, Ricardo Campos, Dimitar Dimitrov, Alípio Jorge, Senja Pollak, Nikolay Ribin, Zoran Fijavž, Maram Hasanain, Nuno Guimarães, Ana Filipa Pacheco, Elisa Sartori, Purificação Silvano, Ana Vitez Zwitter, Ivan Koychev, Nana Yu, Preslav Nakov, and

Giovanni Da San Martino. Overview of the CLEF-2024 CheckThat! Lab Task 3 on Persuasion Techniques. In Faggioli et al. (2024).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, 2018.

L. A. Ramshaw and M. P. Marcus. *Text Chunking Using Transformation-Based Learning*, pages 157–176. Springer Netherlands, Dordrecht, 1999. ISBN 978-94-017-2390-9.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin 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 Lrocessing*, pages 2931–2937, 2017.

Ahmed Sajwani, Alaa El Setohy, Ali Mekky, Diana Turmakhan, Lara Hassan, Mohamed El Zeftawy, Omar El Herraoui, Osama Afzal, Qisheng Liao, Tarek Mahmoud, et al. FRAPPE: FRAming, Persuasion, and Propaganda Explorer. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 207–213, 2024.

Muhammad Salman, Asif Hanif, Shady Shehata, and Preslav Nakov. Detecting Propaganda Techniques in Code-Switched Social Media Text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16794–16812, 2023.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, 2021.

Carol Myers Scotton. The possibility of code-switching: motivation for maintaining multilingualism. *Anthropological linguistics*, pages 432–444, 1982.

Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. Large language models for propaganda detection. *arXiv preprint arXiv:2310.06422*, 2023.

Mary WJ Tay. Code switching and code mixing as a communicative strategy in multilingual discourse. *World Englishes*, 8(3):407–417, 1989.

N. Thurman. Personalization of news. In T. P. Vos, F. Hanusch, D. Dimitrakopoulou, M. Geertsema-Sligh, and A. Sehl, editors, *The International Encyclopedia of Journalism Studies*. Wiley-Blackwell, Massachusetts, USA, May 2018.

José Miguel Túñez-López, Carlos Toural-Bran, and Ana Gabriela Frazão-Nogueira. From data journalism to robotic journalism: The automation of news processing. *Journalistic metamorphosis: media transformation in the digital age*, pages 17–28, 2020.

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

Liqiang Wang, Xiaoyu Shen, Gerard de Melo, and Gerhard Weikum. Cross-domain learning for classifying propaganda in online contents. In *Truth and Trust Online Conference*, pages 21–31. Hacks Hackers, 2020.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Anthony Weston. *A rulebook for arguments*. Hackett Publishing, 2018.

Kai-Cheng Yang, Onur Varol, Clayton A Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61, 2019.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. Unsupervised text style transfer using language models as discriminators. *Advances in Neural Information Processing Systems*, 31, 2018.

Seunghak Yu, Giovanni Da San Martino, Mitra Mohtarami, James Glass, and Preslav Nakov. Interpretable propaganda detection in news articles. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1597–1605, Held Online, September 2021. INCOMA Ltd.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

Jinxue Zhang, Rui Zhang, Yanchao Zhang, and Guanhua Yan. The rise of social botnets: Attacks and countermeasures. *IEEE Transactions on Dependable and Secure Computing*, 15(6):1068–1082, 2016.