Fine-Grained Classification of Polarized and Propagandist Text in News Articles and Political Debates

Vorakit Vorakitphan

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Classification Fine de Textes Polarisés et de Propagande Issus d'Articles d'Actualité et de Débats Politiques

Vorakit VORAKITPHAN

Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis

Présentée en vue de l'obtention du grade de docteur en Informatique **d'**Université Côte d'Azur
Dirigée par : Serena VILLATA
Co-encadrée par : Elena CABRIO
Soutenue le : 15.12.2021

Devant le jury, composé de :
Président du jury : Fabien GANDON, Directeur de recherche, Inria Sophia Antipolis - Méditerranée
Rapporteurs :
Farah BENAMARA, Maîtresse de Conference, IRIT, Université Paul Sabatier
Anne VILNAT, Professeure, LISN, Université Paris-Saclay
Examinateurs : Marco GUERINI, Chercheur, Fondazione Bruno Kessler, Trento, Italy
Laura ALONSO ALEMANY, Professeure, Universidad Nacional de Córdoba, Argentine

Inria i3S sophia antipolis Qwant

Classification Fine de Textes Polarisés et de Propagande Issus d'Articles d'Actualité et de Débats Politiques

Fine-Grained Classification of Polarized and Propagandist Text in News Articles and Political Debates

COMPOSITION DU JURY

Président du jury::

Fabien GANDON, Directeur de recherche, Inria Sophia Antipolis - Méditerranée

Rapporteurs:

Farah BENAMARA, Maîtresse de Conference, IRIT, Université Paul Sabatier, Toulouse
Anne VILNAT, Professeure, LISN, Université Paris-Saclay, Paris

Examinateurs:

Marco GUERINI, Chercheur, Fondazione Bruno Kessler, Trento, Italy
Laura ALONSO ALEMANY, Professeure, Universidad Nacional de Córdoba, Argentina

Directrice de thèse:

Serena VILLATA, Chargée de recherche, CNRS

Encadrante:

Elena CABRIO, Maîtresse de Conference, Université Côte d'Azur

Résumé

La désinformation, dont la propagation s'est accentuée par le biais des médias sociaux, suscite aujourd'hui une réelle menace pour la société. Il existe différents moyens de véhiculer de la désinformation, par exemple par le biais de contenus délibérément manipulés ou fabriqués dans le but de créer des théories conspirationnistes, de rumeurs ou encore de positions et jugements erronés, tels que l'on peut en rencontrer dans des articles d'actualité, discours et débats politiques. L'une des nombreuses formes de désinformation rencontrée en ligne, et certainement l'une des plus dangereuses, est *la propagande*. Ce type de désinformation, que l'on retrouve notamment en politique, représente une stratégie de communication efficace mais souvent trompeuse utilisée pour promouvoir un certain point de vue auprès du public. La nécessité d'identifier, de classifier et de comprendre efficacement et automatiquement ce type de phénomène devient pressant. Dans cette thèse, j'aborde cette question et je propose une approche de classification fine des textes polarisés et de propagande issus d'articles de presse et de débats politiques. Selon le sujet abordé, le contexte, la source d'information, les antécédents et les préférences constituent un panel de facteurs pouvant influer sur les perceptions de l'auditoire et donc conduire à sa déviation ou polarisation en faveur d'un parti. À partir d'un cas d'utilisation provenant d'un scénario politique, nous proposons d'explorer les impacts d'une telle polarisation par le biais de méthodes issues de l'analyse de sentiment basée sur des aspects. L'objectif étant de vérifier dans quelle mesure ces méthodes peuvent permettre d'obtenir des informations sur les messages politiques postés sur les médias sociaux. Plus particulièrement, la thèse traite de la conception et de l'évaluation d'un certain nombre de techniques d'extraction des principales caractéristiques des textes de propagande dans le domaine du Traitement Automatique du Langage Naturel (TALN). L'analyse de sentiment, les techniques de persuasion, la simplicité des messages et l'argumentation y sont

notamment proposées et étudiées en profondeur. Les résultats de cette thèse montrent que ces caractéristiques peuvent capturer des propriétés particulières permettant de caractériser la propagande dans les textes. D'autre part, ces caractéristiques sont employées dans le cadre de la conception et l'implémentation d'une architecture neuronale ayant pour vocation à détecter et classifier les techniques de propagande à grain fin. Le travail proposé dans cette thèse va au-delà de l'état de l'art des systèmes actuels de détection et de classification de la propagande à grain fin. En effet, plusieurs approches d'apprentissage automatique, allant de la régression logistique à des architectures neuronales récentes, ont été testées sur des jeux de données standard de détection de la propagande. En conséquence, un pipeline complet de détection et de classification de la propagande est présenté. La tâche de détection des extraits de textes de propagande a obtenu un score F1 de 0,71, et l'architecture basée sur les transformateurs a obtenu une moyenne de 0,67 pour la tâche de classification des techniques de propagande, surpassant ainsi les systèmes de pointe. Ce pipeline est démontré avec un outil de preuve de concept appelé PROTECT. Enfin, comme dernière contribution de cette thèse, j'ai participé à la création d'une nouvelle ressource linguistique annotée. Composée de textes issus des débats politiques des campagnes présidentielles américaines de 1960 à 2016, cette ressource est annotée avec 6 types de techniques de propagande qui se décomposent en 14 sous-catégories de propagande. L'ensemble de données que j'ai construit contient 1666 instances de propagande.

Mots clés: Traitement Automatique du Langage Naturel, Détection de la Propagande, Analyse de Sentiment, Argumentation

Abstract

In recent years, *disinformation* has become more viral, mainly due to its spread online on social media, leading to potential threatening consequences for the society. Heterogeneous forms of online disinformation are possible, i.e., deliberately manipulated or fabricated content with the intentional aim of creating conspiracy theories, rumours, or misbehaved stances and judgements, for instance, in news articles, and political discourse and debates. One of many instances of online disinformation, and certainly one of the most dangerous ones, is *propaganda*. This disinformation instance represents an effective but often misleading communication strategy which is employed to promote a certain viewpoint to the audience, for instance in the political context. The need to effectively and automatically identify, classify and understand such phenomenon is becoming a urgent need. In this thesis, I tackle this issue and I propose a fine-grained classification approach of polarized and propagandist text in news articles and political debates. More precisely, as the audience' perceptions are perceived differently depending on the context, the source of information, the audience background and preferences, a discussed topic can deviate or polarize the audience into a partisanship. This thesis firstly investigates such polarization given a use-case in a political scenario using Aspects-Based Sentiment Analysis to verify how extensively these methods can be employed to gain insights from the political posts on social media. The thesis discusses the design and evaluation of a number of techniques in extracting the main features of propagandist text in the area of Natural Language Processing (NLP) where sentiment analysis, persuasion techniques, message simplicity, and ultimately argumentation are proposed and thoroughly investigated. The findings in this thesis show that such features can capture particular characteristics of propaganda in texts. Furthermore, these features are employed to tackle the NLP tasks of propaganda detection and classification through the design and implementation

of a neural architecture to classify fine-grained propaganda techniques. The work in this thesis goes beyond the state-of-the-art of current systems for fine-grained propaganda detection and classification. Various Machine Learning approaches ranging from feature-based logistic regression to recent neural architectures have been experimented on standard benchmarks in propaganda detection. As a result, a full pipeline in propaganda detection and classification is presented where the task of detecting the propagandist text snippets obtained .71 F1-score, and the transformer-based architecture obtained an average of .67 F1-score for the task of propaganda technique classification, outperforming the state-of-the-art systems. This pipeline is demonstrated with a proof-of-concept tool called PROTECT. Finally, as a last contribution of this thesis, I carried out the creation of a new annotated linguistic resource. This resource is annotated with 6 types of propaganda techniques, which breaks down into 14 sub-categories of propaganda in the political debates of the US presidential campaigns from 1960 to 2016. The data set I built contains 1666 instances of propagandist text.

Keywords: Natural Language Processing, Propaganda Detection, Sentiment Analysis, Argumentation

Acknowledgements

First and foremost, I am tremendously grateful to Serena and Elena, my PhD supervisors, for their pioneers, masterfulness and hard work to give myself a guidance in scientific and professional prospects. Without them, this journey of my PhD would have been neither uneven nor impossible. I deeply thankful to them for a huge ton of effort, patience and engagements during my PhD study to help sorting out my curiosities into developments in academia.

I would like to express my gratitude to the WIMMICS team, and all members which combines such international, ambitious, and intellectual researchers and students altogether as a team. Thank you, to Fabien who opens up opportunities, advises and great supports. Thank you, to Franck and Marco who encourage my motives during my research toward professional future plans. Thank you, to Christine who always assists helps and provides directions to make my stay in the team well-organized. Thank you, to Amine, Antonia, Tobias and Anna for your academia views, discussions and friendships outside the team. Also, thank you to all members in the team for the insightful and educational coffee-breaks.

My gratitude goes out towards INRIA Sophia Antipolis and QWANT under the ANSWERS project, in addition to the Royal Government of Thailand for providing me with the supports I needed to pursue my studies.

At last, words cannot express enough how thankful I am for my parents and my brother who understand and support my decisions and studies throughout my PhD program.

FUNDING: This work is funded by the ANSWER project PIA FSN2 n. P159564-2661789/DOS0060094 between Inria and Qwant.

Contents

List of Figures

List of Tables

List of Abbreviations

ABSA	Aspect-Based Sentiment Analysis
AI	Artificial Intelligence
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bi-directional Long Short-Term Memory
CPD	Commission on Presidential Debates
CRF	Conditional Random Field
ELMo	Embeddings from Language Model
EU	European Union
FLC	Fragment-Level Classification
GloVe	Global Vectors
IAA	Inter-Annotator Agreement
IOB	Inside–Outside–Beginning
LM	Language Model
LR	Logistic Regression
MSE	Mean Squared Error
NLP	Natural Language Processing

NLTK	Natural Language Toolkit
PLM	Pre-trained Language Model
PoS	Part-of-Speech
PROTECT	PROpaganda Text dEteCTion
Prta	Propaganda Persuasion Techniques Analyzer
ReLU	Rectified Linear Unit
RoBERTa	Robustly Optimized BERT pretraining Approach
RNN	Recurrent Neural Network
SA	Sentiment Analysis
SLC	Sentence-Level Classification
SOTA	State-of-the-Art
SVM	Support Vector Machine
TC	Technique Classification
T5	The Text-To-Text Transformer
UK	United Kingdom
URL	Uniform Resource Locator
US	United States
VAD	Valence-Arousal-Dominance

List of Published Papers

Vorakit Vorakitphan, Marco Guerini, Elena Cabrio, and Serena Villata. Regrexit or not regrexit: Aspect-based sentiment analysis in polarized contexts. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, pages 219–224, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. "Don't discuss": Investigating Semantic and Argumentative Features for Supervised Propagandist Message Detection and Classification. In *Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1498–1507, Varna, Bulgaria (Online), September 2021.

Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. PROTECT: A Pipeline for Propaganda Detection and Classification. In *The Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)*, In press, Milan, Italy, January 2022.

Papers under review

Pierpaolo Goffredo, Shohreh Haddadan, **Vorakit Vorakitphan**, Elena Cabrio, and Serena Villata. Fallacious Argument Classification in Political Debates.

Chapter 1

Introduction

This chapter explains the underlying motivation of the work presented in this thesis. It lays out the context, the main features and the system architecture I proposed to detect and classify propagandist messages. In addition, this section explains the Natural Language Processing methods employed, and evaluation steps taken to address the challenges in the research field of propaganda detection in text. Furthermore, a new linguistic resource annotated with propaganda techniques is presented together with an online tool to identify propagandist text I designed to ensure usability by online users.

1.1 Background and Motivation

Propaganda has mostly served to conceal state-corporate authority from community awareness and engagement. The news media is an apparent route for the spread of propaganda because of its social relevance in the shaping of public opinion. Contemporary communication, media, and journalism studies, on the other hand, have largely overlooked the news media's role in generating and disseminating propaganda. Scholarly documentation on the news media generally employ the word of propaganda [86].

Historically, propaganda is defined as *a weapon to psychologically threaten*

an opponent's esteem [20]. The propaganda phenomenon is used to create and maintain allies and to persuade neutrals to support the action, or remain passive depending on the purpose they were meant to serve. Some studies suggest that propaganda is a subcategory of persuasion which includes a 'hidden agenda' [49].

Propaganda tends to offer ready-made opinions in different contexts, for example in advertising or in education. In the case of a marketing or advertising, the phenomenon is set up aiming to give messages guided by different intentions to persuade audience to agree on the need or demand even if the audience is not in favor of at first. In education, propaganda is seen as a read-made content than has a similar approach to advertising but it aims at delivering the same content that has been agreed by the educational constitutions to be shared across the country or the world which also rely on the independence of the judgment of the audience [13]. In the case of education, students are particularly vulnerable to propaganda, disinformation, and fake news since information and communication technology is fundamental to their being. Young people spend a substantial amount of time online which make them rely significantly on information distributed online. With the rise of fake news as a form of propaganda in recent years, it is more important than ever for students to be able to recognize misleading and biased information.

Propaganda is an effective way to promote a cause or a viewpoint as it consists of psychological shifts and persuasion which can be easily found in politics. There are various communication means that can be applied the spread of disinformation as propaganda. This can be done with textual documents, images, videos and oral speeches. Thus, there is a significant necessity to be able to identify and call out misleading or harmful information in real-world applications.

In modern-day society, the news media serve as a major medium for information distribution. For instance, the news media supply the main percentage of the information on which voters make their political decisions in principle. The notion that the mass media is maintained by large corporate concerns orientated within the social and economic system, contributes to

the system's protection and conservation results through propagandist news media behavior [37; 64]. Many research achievements in political psychology, and computational politics domains [18; 81] suggest that the shift in the structure of social control of disinformation needs more investigation in exploring how to identify propaganda texts. In addition, the COVID-19[1] pandemic has been effecting the world's situation economically, physically and also mentally into reflecting the world's tragedy. Due to this circumstance, the World Health Organization (WHO) undelined that "We're not just fighting a pandemic; we're fighting an infodemic.", meaning that the *misinformation, disinformation, malinformation* are being exhibited heavily in the society.

However, being able to correctly identify propagandist text remains a challenging task. In Natural Language Processing (NLP), the task of detecting propaganda in texts is still relatively new in the domain. Some documents are composed as a long document where propagandist text snippets have to be identified at paragraph-level or sentence-level. In some cases, a post can be short and contain a propagandist text as a span or a snippet. In news articles, the journalist is presumed to be unbiased and topic-oriented, however, this is not the case in some circumstances. Hence, the necessity of being able to detect such propagandist text is risen to an attention of the NLP community. The ultimate goal is to develop active algorithms to learn how to detect such text. Nevertheless, a propaganda text does not consist of solid or well-defined characteristics to be easily identify by using rule-based, even with some traditional machine learning algorithms such as, Naive Bayes classifier, Support Vector Machine (SVM), or Random forests learners. Modern learning algorithms and the State Of The Art (SOTA) models in text processing are concerned to tackle such issues due to a deeper complexity in the computation. Yet, the identification and characteristics of propagandist features are still in question: *how the propaganda text is constructed?, how to extract such features?, and how to employ such features into the SOTA*

[1]Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus.

models?.

Da San Martino et al. [23] demonstrates that the propaganda phenomenon in text is relatively challenging to detect as it combines multiple disciplines related to persuasive techniques in modern mass communication, sentiment analysis, and argumentation mining. Therefore, detecting propaganda information by applying the SOTA language models only can be inadequate as the propaganda phenomenon is highly concerned by complex linguistic phenomena. First, there is the need to understand the characteristics of persuasive text as done in the computational social science and social psychology domains [20]. Dillard and Pfau [29] show that lexical complexity, language intensity, and power of speech style are the most influential characteristics of persuasive text. Longpre et al. [54] investigate what factors are most important in influencing one side to be persuaded. They show that persuasion features can be considered from gender, political and religious ideology, similarity score of opinions, decidedness or undecidedness from audience information. These social science studies should be considered as the baseline to extract the linguistic features associated to propaganda textual content.

Following Bolsover and Howard [14], in this thesis I focus on *computational propaganda*, which is defined as "propaganda created or disseminated using computational (technical) means". Most work in propaganda detection in recent days apply embedding models together with various types of Recurrent Neural Networks (RNN) architectures. However, fine-grained results based on the range from char, token, word to sentence or article levels may differ in overseeing the strategy of building RNN architectures. From a NLP perspective, Da San Martino et al. [21] propose to decompose the Fine-Grained Propaganda Detection task into two sub-tasks:

- *Fragment-Level Classification task (FLC)*: Given a news article, the task is to detect all spans of the text in which a propaganda technique is used. In addition, for each span the applied propaganda technique must be identified.

- *Sentence-Level Classification task (SLC)*: A sentence is considered pro-

pagandist if it contains at least one propagandist fragment. It is usually defined as a binary classification task in which, given a sentence, the correct label, either propaganda or non-propaganda, is to be predicted.

An example of a sentence that converts the labels regarding the subtasks aforementioned: ^{0}*Manchin says Democrats acted like* ^{34}*babies*^{40} *at the SOTU (video) Personal Liberty Poll Exercise your right to vote.* The superscript represents char locations where the word "babies" is considered as "Name_Calling-and-Labeling" technique. I consider this sentence as a propaganda message as there is at least one token (e.g., "babies") is being propagandist in SLC task. On the other hand in FLC task, each token is labeled independently where only "babies" token is labeled as "Name_Calling-and-Labeling" while all other tokens are labeled as "none-propaganda". The thesis goes in the direction of boosting the detection accuracy ranging from SLC, that casts a binary classification problem, to FLC, where both multi-class and multi-label classification problems are concerned.

Hence, the goal of this PhD thesis is to tackle the tasks of propaganda detection and classification in NLP, by designing a system to detect, extract and classify propagandist text, and ultimately aggregate it through a pipeline based on the neural architectures to provide the inquired extracted representation information for analyzing such propagandist text in polarized news articles, and political debates.

1.2 Research Questions

The road map of this PhD thesis consists of multiple stages. More precisely, each stage can be broken down into one research question (RQ) addressing different facets of the problem of detecting and classifying propagandist text:

RQ1: *How basic constituents of emotions and sentiment interact with aspect-based polarization in news articles?* This research question addresses the issue of effectiveness (or not) if the audience is divided or polarized into different groups of interest regarding the disagreement over a certain topic.

The study I address in this thesis is to answer this research question presents how the polarization affects the topic, which can result in some influence over the feature representation. It further concerns the perceived sentiment and emotions, and how the overall output can be aggregated to be applicable in such polarized context.

RQ2: *What are the linguistic distinctive features of propaganda in text and how to tackle the extractions step of such features?* This research question is answered through a detailed linguistics analysis of the features characterising propaganda information in textual format. The question also aims at developing a conceptualization of the results obtained in answering RQ1. The goal is to get more linguistic insights on the components of propagandist text. More precisely, I focus on the propagandist linguistic features related to semantics, sentiment analysis, and argumentation techniques.

RQ3: *What computational approaches can be used to effectively detect and classify propaganda texts in fine-grained settings?* This research question aims at developing a methodology based on the results obtained by answering RQ2. This research question breaks down into the following sub-questions:

- How to integrate the linguistic distinctive features of propaganda text identified in RQ2?
- What are the optimal architectures for automatically detecting propaganda text, given the fine-grained classification task which is target? The goal is to detect the propagandist texts based on different granularity levels of the text (e.g., sentence-level, token-level, fragment-level). This comprises challenges in adapting multiple implementation architectures based on specific granularity levels, as well as the pre- and post-processing steps to fit the settings of each text granularity as the length of each propaganda snippet can be varied.
- How to identify the different propaganda techniques according to the standard fine-grained categories? This consists of determining how to implement the detection of the propaganda text, then further classify the techniques applied.

1.3 Contributions

The main contributions of this thesis are as follows:

Contribution 1 - Polarization in Political News Articles and Investigation of Linguistic Distinctive Features of Propaganda. The first contribution of this PhD thesis aims to address RQ1 and RQ2. First, to address RQ1, the context of news articles and political posts are analysed to validate the consistency between a topic and the audience. A single discussed topic can deviate the audience into different aspects regarding their agreements. I studied polarization in order to get a better understanding of how the aspects can be analysed in terms of polarization and partisanship toward stereotypical aspect-based sentiment analysis. A real political use case scenario, i.e., *"the Brexit"*, is chosen, along with the news articles to be examined. In the context of this thesis, I investigated sentiment analysis in polarized contexts, with a special focus on the sentiment, emotions and Valence-Arousal-Dominance (VAD) on the Brexit use case. These understandings are then employed to address propagandist text feature extraction before tackling the propaganda detection and classification tasks. In particular, the distinctive linguistic features of propaganda are extracted and evaluated to answer RQ2. Apart from sentiment features, that can potentially shelter the polarization in news articles and political posts, the contribution in answering RQ2 aims to explore the most influential features in propaganda. Features in the area of persuasion, simplicity of message and argumentation are proposed and assessed in the light of the hypothesis that such features can boost the results of the propaganda detection and classification tasks.

Related Publications

1. **Vorakit Vorakitphan**, Marco Guerini, Elena Cabrio, and Serena Villata. Regrexit or not regrexit: Aspect-based sentiment analysis in polarized contexts. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, pages 219–224, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics.

2. **Vorakit Vorakitphan**, Elena Cabrio, and Serena Villata. "Don't discuss": Investigating Semantic and Argumentative Features for Supervised Propagandist Message Detection and Classification. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1498–1507, Varna (Online), Bulgaria, September 2021.

Contribution 2 - Fine-grained Architectures for the Identification and Classification of Propagandist Texts. After being able to identify distinctive features of propaganda texts, the follow-up contribution goes in the direction of answering RQ3 in proposing new NLP neural architectures to help detecting and classifying propagandist text, and ultimately boost the performance for these tasks. This contribution of the thesis addresses the fine-grained settings of propaganda detection and classification as a pipeline. Such fine-grained settings consist in the propaganda technique classification task cast ranging from sentence-, to token-, to fragment-level classification. To address these tasks I applied different approaches ranging from feature-based logistic regression to Recurrent Neural Networks and other neural methods. In the end, the best performing method relies on an neural architecture for sentence-span classification employing transformer-based models combined with the propaganda distinctive features identified in Contribution 1. Ultimately, a complete pipeline for processing a propagandist text detection and classification is proposed, called PROTECT, addressing *(i)* a propaganda snippet detection returning all the propagandist text snippets extracted from a given text, *(ii)* a fine-grained propaganda technique classification, and *(iii)* the generation of a downloadable *json* file with the annotated text with the propaganda techniques.

Related Publications

1. **Vorakit Vorakitphan**, Elena Cabrio, and Serena Villata. "Don't discuss": Investigating Semantic and Argumentative Features for Supervised Propagandist Message Detection and Classification. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1498–1507, Varna (Online), Bulgaria, September 2021.

2. **Vorakit Vorakitphan**, Elena Cabrio, and Serena Villata. PROTECT: A Pipeline for Propaganda Detection and Classification. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)*, Accepted, in press, Milan, Italy, January 2022.

Contribution 3 - The Creation of an Annotated Linguistic Resource of Propaganda Techniques in Political Debates. The final contribution of this thesis answers RQ3 in detecting propaganda in the political scenario, and closes the gap returning back to the initial use case from RQ1. I moved from the exploration of propaganda in political posts to political debates, where propagandist text is more complex to identify due to the political discourse structure and the complexity of the addressed topics. To the best of my knowledge, there is no other available dataset where political debates are annotated with propaganda techniques. The new dataset I built contains all the transcripts from the United States political debates for the presidential elections from 1960 to 2016. This data set is annotated with 6 categories of propaganda techniques, and 14 sub-categories, leading to a total of 1666 instances of propagandist text.

Ongoing Publications

- We plan to submit this contribution at the beginning of January 2022.

1.4 Structure

The thesis is organized as follows:

Chapter 2 firstly discusses the more frequent use case scenarios where propaganda may arise, i.e., in polarized context in news articles, and political discussions. It outlines the resources built for the fine-grained tasks of propaganda detection and classification, reviews the main standard baselines in the propaganda classification architectures and the methods employed, along with natural language representations of the features.

Chapter 3 firstly addresses the consistency of the relations between the topic of a news and its audience, and secondly discusses the features based on sentiment analysis to be employed in the context of propaganda detection. This chapter presents an analysis of political news articles where the context is polarized. The proposed pipeline applies aspects-based sentiment analysis techniques to characterize polarization.

Chapter 4 investigates the fine-grained propaganda classification tasks and highlights the influential features in propagandist text detection and classification. Features in semantics, persuasion, sentiment analysis, message simplicity, and argumentation are examined with the final aim of boosting the performance in the detection and classification of propaganda in fine-grained settings from sentence- to fragment-level. An in-depth error analysis of the proposed architectures is addressed.

Chapter 5 describes the new dataset of US presidential debates I built annotated with the propaganda techniques. This chapter details the whole annotation process, from the definition of the annotation guidelines to the annotation platform, with the help of examples from the debates and exceptions to the annotation guidelines. Lastly, the evaluation and data statistics of this new linguistic resource are presented.

Chapter 6 presents a full pipeline of propaganda detection and classification implemented into a proof-of-concept online tool called PROTECT. The full pipeline allows the user to automatically analyse their text, with the aim to identify the propagandist text snippets in there, if any. The result is then returned to the user under the form of annotated text, and word clouds based on the given text.

Chapter 7 summarizes the main contributions of the thesis and the remaining open challenges. In addition, prospective applications and future research directions are discussed.

Chapter 2

Related Work

This chapter summarizes and discusses the relevant research in the area of fine-grained propaganda detection and classification, and the feature representation for these tasks. It furthermore brings the findings in this thesis into context by comparing with state-of-the-art systems for propaganda detection and fine-grained classification in the context of polarized contexts in news articles and political debates to underline differences and commonalities among existing approaches.

This chapter presents existing resources in propagandist text, and highlights various approaches to automatically detect and classify propaganda. In particular, Section 2.1 introduces various textual sources which can be employed for these tasks, especially in political posts and polarized contexts. Section 2.2 discusses the tasks of propaganda detection and classification in text. Section 2.2.1 points out the existing textual resources annotated for the propaganda detection and classification tasks. Section 2.2.2 investigates the relevant factors that potentially influence propagandist texts. Section 2.2.3 describes current classification architectures in propaganda detection. Section 2.2.4 discusses the difficulty in evaluating the fine-grained settings of propaganda texts. Section 2.2.5 presents current techniques in embedding and feature representation toward the use of RNNs and transformer-based models. Lastly, the list of applications in propaganda detection is provided

in Section 2.3.

2.1 Propaganda in Online Posts

Nowadays, propaganda appears online oftentimes in form of text, and is considered as "digital misinformation and manipulation efforts (disinformation)" Woolley and Howard [83]. The most common means in delivering such propagandist texts are via social media and news articles. However, verbal mean of persuading and misleading through a message is still valid and mostly used in political debates [35], for instance. Most of the time, propaganda has been considered by many researchers as misleading information [83; 49; 72], leading to an increased interest toward *computational propaganda* in real-world applications. The study of propaganda in the context of this thesis focuses mainly on news articles and political discussions. Below I explain in which context propaganda is usually employed.

News Articles. Propaganda techniques are used in persuasion discreetly and psychologically to convey the audience. Toward such goals, news articles, where people are most influenced onto, usually contain propaganda techniques. Habitually, a propaganda content is approached as a reliable content provided, however, without the adequate transparency about the source of the news and the motivation. The purpose of news propaganda can be to enhance or maintain government, political, or ideological motives, partisan agendas, religious or ethnic motives, or business or commercial motives. Hence, to determine the content to be propagandist or traditional press releases, the transparency of the news source usually is one of the main parameters. In [21; 22], the authors propose the task in detecting propagandist texts on political news articles. The work conducted starts from an annotation task of 14 to 18 propaganda techniques on news articles until the automatic detection tasks. The authors refer to a propaganda message in news articles in the form of snippets within a post.

Political Discussions. Some research focus heavily on the involvement of propaganda techniques in texts among online communities in the domain

of politics based on different sorts of news events [63]. Meanwhile, another direction in computational propaganda addresses the relationship in political communities on the Reddit[1] website based on the submitted posts along with the distribution of the received attention [68]. The work of Balalau and Horincar [7] also focuses on political issues discussed in online forums, by proposing a neural architecture to assess how propagandist texts can diversify the audience. Their study covers political discussions on two use cases (e.g., the United States and the United Kingdom) to identify which group of minorities tends to use more propaganda and contain more political biases. The employment of different propaganda techniques from the left-wing and the right-wing communities of each use case are also investigated. Carman et al. [17] study the effect of vote manipulation and user engagement based on the reader voting mechanisms on Reddit, to assess if such factors can promote or suppress the visibility of articles and reviews in political threads. An et al. [3] examine an analytical guide in political discussions to investigate the characteristics, interaction and linguistic patterns between heterogeneous communication areas and politically homogenized Reddit posts. Furthermore, Guimaraes et al. [43] analyze characteristics of user actions (i.e., replies and votes) and identify the different patterns that re-fine the notion of controversies into disputes, disruptions and discrepancies in political threads on Reddit. This work also clarifies how sentiment can be engaged in posts and replies within such political discussions.

Polarized Context in Politics. Demszky et al. [27] apply Aspect-Based Sentiment Analysis (ABSA) methods to analyse tweets on the US mass shootings topic, where the topic was politically discussed from different viewpoints according to the locations of the event, with the contrasting use of the terms “terrorist” and “crazy”, that contribute to polarization. There are the further understandings addressed by Demszky et al. [27] such as a discussion about how such concepts are evaluated – in polarized scenarios – by both parties (e.g., how people and journalists from each stance react when the same shooter is defined as “terrorist” or “crazy”). Such a polarized

[1]https://www.reddit.com/

context uses different techniques in propaganda to convey their audience, even though all sources are talking about the same topic. In consequence, ABSA approaches aim not only at capturing linguistic relations, but also at ascribing polarity stances of textual entities. To some extent, the existing works in detecting textual elements in news article can expand toward the detection of fake news where the media bias can politically occur. Concerning such direction, Iyyer et al. [46] proposed a framework taking into account the technique in sentiment analysis to conduct RNN based automated techniques to identify ideology from text focusing political position evinced by a sentence. Toward the political stance detection and fact-checking direction, Baly et al. [9] explored political ideology or bias detection on news articles where partisanship is split into left, center, or right. The author proposed an adversarial media adaptation where background information about the source at article-level to prevent modeling the source instead of the political bias in the news articles. Later on, Stefanov et al. [70] conducted a framework on Twitter data concerning a user stance detection and predicting media bias. The framework consists of an unsupervised method to ascertain stance based on polarized topics regarding retweet behavior, then a supervised classification method to characterize both the general political leaning of online media and of popular Twitter users.

Other NLP approaches show that certain topics are more polarizing than others (Balasubramanyan et al. [8]). In the context of this thesis, my work [75] explains a discussion topic in political news articles which diversify the aspects of audience into several groups and can be seen as a polarized context. The topic of "Brexit"[2] was selected as use case where the news sources are discussed to understand whether they were *supporting* or *against* the concept of Brexit. The work examines how elements of sentiment, emotion and Valence-Arousal-Dominance (VAD) in NLP are correlated in the polarized context. At this point, an Aspect-based Sentiment Analysis approach is applied based on the shared Brexit topic that contains opposite aspects for the audience. This study shows that using a standard sentiment analysis method

[2]The withdrawal of the United Kingdom from the European Union.

on a (polarized) topic across various news articles can suppress the polarized sentiment, and as a consequence, the conclusion is misleading.

Hence, sentiment analysis can potentially boost the accuracy in determining sentiment obtained in such circumstance of aspects on a topic toward propaganda. In contrast, the ABSA method consists in an unsupervised approach where aspects are conveyed with respect to fine-grained extraction of affects in polarized situations. This thesis highlights how sentiment analysis methods help guiding the detection of propaganda texts both in traditional news articles (single aspect-based), and in polarized context.

2.2 Detection and Classification of Propaganda

In this section, I present the tasks of propaganda detection and classification, with a focus on the existing textual resources annotated for these tasks, the relevant factors that influence propagandist text detection, the current classification architectures, and the evaluation of such approaches.

2.2.1 Resources in Propaganda Texts

Resource	Granularity	Annotation	#Articles	#Classes	All classes
Rashkin et al. [61]	Document	News sources	22,580	4	Satire, Hoax, Propaganda, Reliable News
Barrón-Cedeño et al. [10]	Document	Max. entropy	51,294	2	Propaganda, Trustworthy
NLP4IF'19 Da San Martino et al. [21]	Text span (Fragment and sentence)	Professional annotators	451	18	Loaded_language, Name_calling labeling, Repetition, Exaggeration minimiz., Doubt, Appeal_to_fear prejudice, Flag-waving, Causal_oversimplification, Slogans, Appeal_to_authority, Black-and-white_fallacy, Thought-Bandwagon, terminating_cliches, Whataboutism, Reductio_ad_hitlerum, Red_herring, Obfuscation intentional vagueness confusion, Straw_man
SemEval'20 Task 11 Da San Martino et al. [22]	Text span (Fragment and sentence)	Professional annotators	539	14	Loaded_language, Name_calling labeling, Repetition, minimizat., Doubt, Appeal_to_fear prejudice, Flag-Waving, Causal_oversimplification, Slogans, Appeal_to_Authority, Black-and-White_fallacy, Thought-terminating_cliches, Whataboutism straw_men red_Herring, Bandwagon Reductio_ad_Hitlerum

Table 2.1: Available linguistic resources with propaganda-related annotations.

Given that propaganda texts show a high linguistic complexity, a fine-grained propaganda classification task remains challenging.

In the last years, there has been an increasing interest in investigating methods for textual propaganda detection and classification. Among them, Barrón-Cedeño et al. [10] present a system to organize news events according to the level of propagandist content in the articles, and introduces a new corpus (QProp) annotated with the propaganda vs. trustworthy classes, providing information about the source of the news articles. Da San Martino et al. [21] present the benchmark of the shared task NLP4IF'19[3] on fine-grained propaganda detection. The training, the development, and the test partitions of the corpus used for the shared task consist of 350, 61, and 86 articles and of 16,965, 2,235, and 3,526 sentences, respectively. The task has a fine-grained setting in identifying propaganda into two sub-tasks: *i) Sentence-Level Classification task (SLC)*, which asks to predict whether a sentence contains at least one propaganda technique where a sentence is considered propagandist if it contains at least one propagandist fragment, and *ii) Fragment-Level Classification task (FLC)*, which asks to identify both the spans and the type of propaganda technique. As a follow up, in 2020 SemEval proposed a shared task (T11) (Da San Martino et al. [22]) reducing the number of propaganda categories with respect to NLP4IF'19, and proposing a more restrictive evaluation scheme where the sub-tasks are divided into two: *i)* Span Identification task (SI) where a plain-text document is given aiming to identify those specific fragments that contain at least one propaganda technique, and *ii)* Technique Classification task (TC) where a propagandistic text snippet is given along with its document context aim at identifying the propaganda technique used in that snippet. The SemEval2020 T11 contains the training, development and test sets of 371, 75, and 90 articles, respectively. Table 2.1 reports on the available resources annotated for propaganda detection and classification tasks at different granularity levels.

Given the scarcity of annotated resources for this task, one of the main contributions in this thesis is the annotation of a new linguistic resource of political debates annotated with 6 fine-grained propaganda techniques, and 14 sub-categories. Chapter 5 describes the annotation guidelines and reports

[3] `https://propaganda.qcri.org/nlp4if-shared-task/`

on the main statistics of the collected data set. This annotated resource would be helpful to the community to build more effective propaganda detection and classification tools and boost the results on these tasks.

2.2.2 Feature Representation in Propagandist Text

Argumentation. Extracting argument components from propaganda text is the main focus of Durmus and Cardie [33]). Their work consists in analyzing the characteristics of argumentative text in order distinguish the persuasive propagandist text from non-persuasive one e.g., tf-idf scores for unigrams and bigrams, ratio of quotation marks, exclamation marks, modal verbs, stop words, type-token ratio, hedging, named entity types, POS n-grams, sentiment, and subjectivity scores, spell-checking, argument lexicon features (see also Durmus et al. [34] for a detailed discussion of these features).

Sentiment Analysis. Undoubtedly, propaganda text involves concept of sentiment to manipulate a hidden agenda. Travis [72] introduces how propaganda text constructs emotional markers and affect on words or phrases based on lexicon and ontology usage. The work of Ahmad et al. [2] has shown that most recent and effective techniques used nowadays on feature extraction for sentiment in propaganda detection are lexicon-based, context-based tailored dictionaries.

Simplicity of the Message. Propagandist text uses simple terms to convey and ensure that the target perceives the intention successfully. Exaggeration can play a significant role to overact an actual meaning of a word. Li et al. [52] show how to detect different levels in the strength of calmness toward exaggeration in press releases. Additionally, Troiano et al. [73] focus on features extraction for text exaggeration and show that the main factors include imageability, unexpectedness, and the polarity of a sentence.

2.2.3 Classification Architectures

The most recent approaches for propaganda detection are based on language models that mostly involve transformer-based architectures. The approach

that performed best on the NLP4IF'19 sentence-level classification task relies on the BERT architecture with hyperparameters tuning without activation function by Mapes et al. [55]. Yoosuf and Yang [84] focused first on the preprocessing steps to provide more information regarding the language model along with existing propaganda techniques, then they employ the BERT architecture casting the task as a sequence labeling problem. Such detection tasks are fine-grained tasks that cast the granularity of the propagandist text snippet ranging from sentence-, token-, and fragment- (span) level. The NLP4IF'19 challenge applies the objective of sentence-, and token- level classification. In the SemEval 2020 Challenge - Task 11, the systems that took part are the most recent approaches to identify propaganda techniques based on given propagandist spans. This challenge is based on fragment-level classification approach that crucially involves multiple (stacked) transformer-based architectures which are exceedingly computational expensive. The most successful approach is the one proposed by Jurkiewicz et al. [47] where they first extend the training data of the challenge from a free text corpus as a silver data set, and second, they propose an ensemble model that exploits both the gold and silver data sets during the training steps to achieve the highest scores. Despite the effectiveness of these computational architectures, propagandist messages require also the understanding about impact of the linguistic features.

In this thesis, the language model architectures are the principle toward the building of the detection and classification of propaganda messages, empowering them with a rich set of features that I identified as pivotal in propagandist text from the computational social science literature. In particular, Travis [72] discusses how emotional markers and affect at word- and phrase-level are employed in propaganda text, whilst Ahmad et al. [2] show that the most effective technique to extract sentiment for the propaganda detection task is to rely on lexicon-based tailored dictionaries. Recent studies, Li et al. [52] show how to detect degrees of strength from calmness to exaggeration in press releases. Finally, Troiano et al. [73] focus on the feature extraction of textual exaggeration and show that imageability, unexpectedness, and the polarity of a sentence are the main factors.

2.2.4 Evaluation of Propaganda Detection and Classification Systems

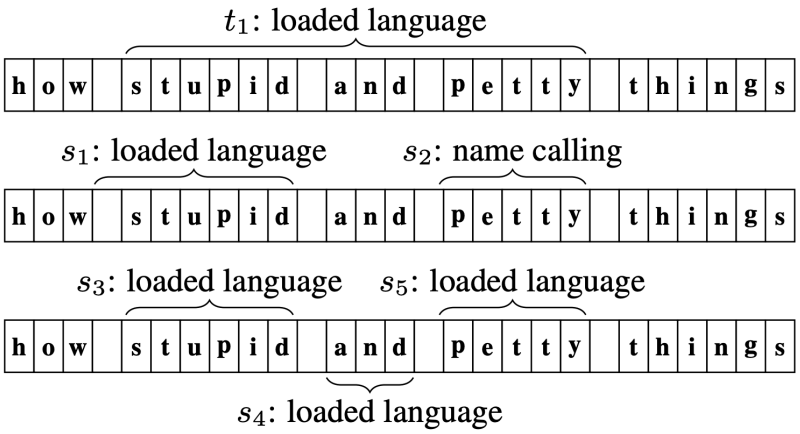


Figure 2.1: Example of gold annotation of a propaganda snippet t, and the snippet predictions s in a document represented as a sequence of characters addressed by Da San Martino et al. [21].

Given a propaganda detection system ranging from detecting at sentence to fragment-level in an article, the evaluation is done differently. In the evaluation at fragment-level, some chars can be overlapped, or dissociated as snippets when compared the prediction and the gold standard. Any of the circumstance in detecting such snippets would effect the score of the matching snippets. For an evaluation at sentence-level, most of the systems cast it as a binary task. Concerning the task of evaluating snippets at fragment-level, the Fragment-Level Classification Task (FLC) by Da San Martino et al. [21] propose to perform the evaluation that considers at char-level where they account for a partial matching between the snippets. Figure 2.1 shows the gold annotation of a propaganda snippet compared against the predicted snippets where each predicted snippet s is predicted per token.

In contrast, the Technique Classification task (TC) by Da San Martino et al. [22] casts the fragment-level evaluation based on a gold standard provided as a snippet-template. This method concerns the multi-class task where the evaluation scores take into consideration only the predicted class

per text, regardless the overlapped, or dissociated char in each snippet. Comparing these two methods in evaluating at fragment-level can lead to different classification scores depending on the classification strategy.

2.2.5 Natural Language Representation for Propaganda Detection

In NLP, word embeddings represent a distributed representation that is learned based on the usage of words. This allows words that are used in similar ways to result in similar representations, naturally capturing therefore their meaning [66; 67; 32]. An embedding captures some of the semantics of the input text by placing semantically similar inputs close together in the embedding space. An embedding can be learned and reused across models. This approach is very popular in current NLP tasks as the feature representation [50; 41].

Toward automated detection of propaganda messages, current systems address these tasks relying on word embedding models (e.g., GloVe (Pennington et al. [58]), Word2Vec (Mikolov et al. [56]), ELMo (Peters et al. [59])) as feature representations to feed various RNN architectures [10; 61]. Recently, the state-of-the-art language models utilize the pre-trained approach along with an attention mechanism of character positions of the text elements to construct the models where vectors are embedded and obtained by various corpus of texts. Such pre-trained models then apply to Recurrent Neural Networks (RNN) architectures (i.e., transformers). This approach has been widely employed to optimize the performances of these classification tasks. One of the most well-known transformer is BERT (Devlin et al. [28]). However, using BERT language model as SOTA, there is still has room to improve the performance of the detection of propaganda in textual messages [22; 21]. For a fine-grained classification of the propaganda techniques, it is a current challenge in recent studies of propaganda detection the fragment-level classification task along with the boundary detection and classification at token-level. Besides, to help coping with the challenge in detection and classification of fine-grained propaganda messages, there are significant factors

that can critically help boosting the accuracy. Firstly, the adaptation of complex computational layers using SOTA classification architectures. Secondly, the language models with engineered features that could potentially extract useful characteristics of propaganda messages. This is precisely one of the main objectives and contributions of this thesis.

2.3 Applications of Propaganda Detection Tools

In the last years, there has been an increasing interest in investigating methods for textual propaganda detection and classification. Among them, Barrón-Cedeño et al. [10] present a system to organize news events according to the level of propagandist content in the articles, and introduce a new corpus (QProp) annotated with the propaganda vs. trustworthy classes, providing information about the source of the news articles. Recently, a web demo named `Prta` has been proposed by Da San Martino et al. [24], trained on disinformation articles. The demo allows a user to enter a plain text or a URL but it does not allow users to download such results. The systems shows the propagandist messages at the snippet level with an option to filter propaganda technique to present based on its confidence rate, and also analyzes the usage of propaganda technique on determined topics. The implementation of this system relies on the approach proposed by Da San Martino et al. [21].

In this line, and with the goal of proposing a proof-of-concept of our propaganda detection and classification pipeline online for the use of a wide public, Chapter 6 presents the PROTECT (A Pipeline for Propaganda Detection and Classification) tool. This propaganda detection tool implements a pipeline to perform the detection at fine-grained levels of propaganda techniques and allows users to download their results under the form of an automatically annotated text with propaganda techniques. PROTECT relies on language model architectures for the detection and classification of propaganda messages, empowering them with a rich set of features I identified as pivotal in propagandist text from the computational social science literature

presented in Chapter 4. In particular, Travis [72] discusses how emotional markers and affect at word- or phrase-level are employed in propaganda text, whilst Ahmad et al. [2] show that the most effective technique to extract sentiment for the propaganda detection task is to rely on lexicon-based tailored dictionaries. Moreover, Li et al. [52] show how to detect degrees of strength from calmness to exaggeration in press releases. Finally, Troiano et al. [73] focus on feature extraction of text exaggeration and show that main factors include imageability, unexpectedness, and the polarity of a sentence. The PROTECT tool adopts a supervised approach based on RNN and pre-trained models to classify textual snippets both as propaganda messages and according to the precise applied propaganda technique, employing a detailed linguistic analysis of the features characterising propaganda information in text (e.g., semantic, sentiment and argumentation features) to get deeper insights on the results obtained by the tool. In addition, the user-friendly interface allows to visualize the word clouds as a groundwork toward the interpretation of the propagandist text snippets identified by the tool. The two services are trained based on two standard benchmarks, namely NLP4IF'2019, and SemEval2020 Task-11.

Chapter 3

Context Polarization Toward Sentiment and Emotions

This chapter investigates the role of sentiment and emotions in context polarization, which plays a key role in propaganda. As propaganda often appears in news articles, this chapter focuses on the study of the interrelation of sentiment and emotional elements in such a way that they can be profitably employed in the propaganda detection task. Emotion analysis in polarized contexts represents a challenge in Natural Language Processing. As a step in the aforementioned direction, I present a methodology to extend the task of Aspect-based Sentiment Analysis (ABSA) towards the affect and emotion representation in polarized settings. In particular, I adopt the three-dimensional model of affect based on Valence, Arousal, and Dominance (VAD). I then present a Brexit scenario that proves how affect varies toward the same aspect when politically polarized stances are presented. The approach captures aspect-based polarization from newspapers regarding the Brexit scenario of 1.2m entities at sentence-level. This chapter demonstrates how basic constituents of emotions can be mapped to the VAD model, along with their interactions in polarized contexts with ABSA settings using biased key-concepts (e.g., "stop

Brexit" vs. "support Brexit"). Quite intriguingly, the framework achieves to produce coherent aspect evidences of Brexit's stance from key-concepts, showing that VAD influence the support and opposition aspects. This chapter presents the results published at the 28th International Conference on Computational Linguistics (COLING-2020) [75].

This chapter analyses the role of emotions and sentiment in polarized news articles. This is a first but mandatory step towards a better understanding of the propaganda phenomenon in news articles, and, more precisely, it helps in achieving a better comprehension on the role of emotions and sentiment as features in propaganda detection neural architectures. Da San Martino et al. [23] suggested that one of the main approaches which allows to identify a propagandist text is sentiment analysis.

Aspect-based Sentiment Analysis (ABSA) aims at capturing sentiment (i.e., positive, negative or neutral) expressed toward each aspect (i.e., attribute) of a target entity. The main interest is to capture sentiment nuances about different entities. However, in a context of opinion polarization, different groups of people can form strong convictions of competing opinions on such target entities, resulting in different (often opposite) evaluations of the same aspect. Compare, for example, the differences in the pro- and anti-Brexit discourses concerning the withdrawal of the United Kingdom from the European Union (EU), aligning with contrasting attitudes toward the EU, the immigration and the country's culture. In fact, while in standard scenarios of sentiment analysis about specific entities and their aspects it is assumed that sentiment is consistent (e.g., *a big screen is a desirable characteristics for a TV*), this is not the case for polarized contexts. Hence, for example, a "clean Brexit" might be desirable to some, but not to others.

Some previous work has showed that sentiment alone (i.e., positive, negative, neutral) is not able to grasp complex emotion phenomena in information diffusion scenarios [42; 12]. Following the suggestion of Guerini and Staiano [42], to analyse in a fine-grained manner our scenario where the same topic can appear in different news outlets that have opposed opinions, I decide to

make resort to the model presented in [65; 15]. These works developed a circumplex model, with a three dimensional space of affect along the Valence, Arousal, and Dominance coordinates, where all emotions can be mapped in the same dimensional space. I therefore hypothesize that VAD plays a significant role in the interaction with linguistic elements in ABSA settings where polarized opinions are at play.

More specifically, I answer the following research questions: *how basic constituents of emotions, such as VAD, interact with aspect-based polarization? How polarized settings differ from standard SA scenarios? How aspects of Brexit and related key-concepts are perceived by opposite parties?*

To answer these research questions, I propose a comprehensive framework for studying the interaction of ABSA with opinion polarization in newspapers and social media. I first trained an RNN that detects the emotions and their intensities at sentence-level, and then I map emotion intensities into the VAD circumplex model. Later, I build a framework to assess whether and how VAD are connected to polarized contexts, by computing the VAD scores of a set of key-concepts that can be found on newspapers with opposite views. These key-concepts (e.g., "stop immigration") are built from a set of aspects ("immigration" in our example) combined with relevant verbs or adjectives that represent a clear polarized opinion toward the aspect (e.g., "stop").

To experiment with the proposed approach, I focus on the Brexit scenario, whose unique circumstances have boosted opinion polarization because of the extremely rapid and volatile changes in the political panorama as the results of the EU referendum were announced (political leaders were quick to say that the outcome had "revealed a divided Britain")[1]. While an analysis of this situation from the viewpoint of political sciences is not the focus of this chapter, the Brexit scenario as discussed on newspapers and on social media provides me with the required elements to carry out my study, because of the opinion divisions formed around one or more political positions or issues. Moreover, to be meaningfully polarizing, issues need to be important (or "salient") to a large section of the public, not just a minority of people with

[1] `https://www.kcl.ac.uk/policy-institute/assets/divided-britain.pdf`

strongly held views, and in the Brexit scenario this is actually the case. In this experimental setting, I select two British newspapers known to be polarized, i.e., either for or against Brexit. Results show that VAD are not absolute, but relative to the newspaper's viewpoint on the key-concept. My approach highlights that using the proposed key-concepts gives us fine-grained details about VAD elements that strongly interact with the polarized context. I show that standard SA approaches can be deceptive in such polarized setting (considering only the word "Brexit" on both newspapers, the valence is almost identical), while my ABSA approach shows a clear-cut polarization.

The structure of this chapter is as follows: Section 3.1 explains the methods applied to collect the data and build the dataset annotated with emotions, Section 3.2 discusses the sentiment elements and a proposed scheme to transform emotions into sentiment. Finally, Section 3.3 presents the *Brexit* use case to perform an analysis of the polarized context on a proposed scheme of sentiment and emotions, and Section 3.4 discusses the main results presented in the chapter and future research directions on polarized contexts and their role in propaganda.

3.1 Emotions Analysis

As the first step of the proposed framework, I focus on the task of detecting emotions and their intensities at sentence-level from the news context. I compare two standard emotion recognition approaches, i.e., a lexicon-based model called DepecheMood++ proposed by Araque et al. [4], and an RNN based approach that casts the task of multi-label ABSA as a regression problem. Both DepecheMood++ and the RNN model were built and trained on the same set of news extracted from the *Rappler.com* website. *Rappler.com* is an online news publisher, where each article is associated to a mood meter. Such mood meter allows each reader to vote the emotions evoked by the article after reading. I implemented a web crawler and harvested a total of 67,828 articles from January 2017 until April 2019 along with the 8 possible emotions voted by readers (i.e., *happy, inspired, amused, afraid, an-*

Emotions	DepecheMood++	LSTM	BiLSTM	CNN
ANGER	0.47	0.62	**0.67**	0.64
FEAR	0.60	**0.78**	0.76	0.73
JOY	0.38	**0.64**	0.63	0.51
SADNESS	0.46	**0.75**	**0.75**	0.7
SURPRISE	0.21	0.43	**0.52**	0.48
AVG All Emotions	0.43	0.64	**0.66**	0.61

Table 3.1: Pearson's correlation (r) on Cross-dataset Testing of Emotion Recognition Models.

gry, annoyed, don't care, sad). As this chapter aims at detecting emotions at sentence-level, I pick all headlines of such articles and the associated emotions as my training set. For the RNN based approach, I use a deep contextualized word representation as my features, namely ELMo (Peters et al. [59]), with a deep bidirectional language model (BiLSTM). I use BiLSTM of 128 dimensional layers with a Rectified Linear Unit (ReLU) activation function (Agarap [1]). Then, I apply Mean Squared Error (MSE) as a loss function with a standard adam optimizer. At the last layer, I use sigmoid function to gate our multi-label output neurons as the intensity of each emotion. To validate the model, I perform cross-dataset testing on the "Affective Text" dataset from SemEval2007-T14 published by Strapparava and Mihalcea [71] using Pearson's correlation (r) to evaluate the performances of the system on each individual emotion. In Table 3.1[2], I report the results for each emotion, and the average of all emotions produced by the different models. Given that the ELMo embeddings with BiLSTM model outperformed the other models, I selected the output of this model as input for the next step (i.e., to map the predicted emotions to the VAD model).

[2]The "DISGUST" label is excluded as there is no correspondence to Rappler emotions.

3.2 VAD Analysis

The VAD model of affects [65; 15] has been extensively used, and relies on three dimensions: *Valence*, that can be positive/negative (e.g., FEAR has a negative valence, JOY has a positive valence) and corresponds to the standard dimension of sentiment analysis; *Arousal*, that can be low or high and ranges from "calm" to "excitement" (e.g., SADNESS has low arousal while ANGER has high arousal, even if they have the same sentiment); finally, *Dominance*, that can be low or high as well, and ranges from "controlled" to "in control" (e.g. FEAR refers to low dominance, INSPIRED to high dominance).

To map the emotions expressed in a sentence to the VAD model, I followed the approach of Guerini and Staiano [42]. I use the VAD scores of emotion labels provided by Warriner et al. [80] which serves as our gold standard to obtain VAD scores from emotion intensities. In particular, here below I report the procedure for Valence computation (Arousal and Dominance computation are akin): given a sentence S, to obtain its valence S_v I multiply the intensity of each emotion $I(e_i)$ – computed using the approach described in Section 3.1 – with the corresponding Valence score $V(e_i)$ present by Warriner et al. [80]. Then, I sum the n emotional dimensions to obtain the final Valence score, as expressed in Equation 3.1:

$$S_v = \sum_{i=1}^{n} V(e_i) \times I(e_i) \tag{3.1}$$

After the conversion, VAD values are in absolute form (arbitrary range). Since newspapers could have different "affective styles" – that are prior to their stance on a specific topic – I apply standardization individually to each VAD dimension for each newspaper to make results better comparable across newspapers.

3.3 A Polarized Context Scenario: the Brexit

Dataset collection. To clearly define the use case, I conducted a survey by asking to 10 British participants currently living in England to agree on the stance of UK newspapers with respect to the Brexit, i.e., they were asked which newspapers are known to be pro- or anti- Brexit. As a result, I selected two newspapers that obtained the mutual agreement across all participants, i.e., The Sun (*thesun.co.uk*) as the pro-Brexit newspaper, and The Guardian (*theguardian.com*) as the anti-Brexit one. I built a web crawler to harvest the articles containing the word "brexit" from these two news websites from year 2017 to 2019. I obtained 28,212 articles in total, and segmented them at sentence-level resulting in 1.2 million sentences.

Key-concepts. For the analysis, I focused on a set of key-concepts relevant to the Brexit discussion (e.g., "brexit", "immigration", "EU"). I created two lists of words that could potentially trigger polarization occurring with the identified key-concepts in text (e.g., "stop", "promote"). These words have then been combined with each key-concept to obtain the "support" and "against" aspect clusters, e.g., "stop Brexit", "block Brexit" for the against aspect, and "support Brexit", "make Brexit" for the support aspect. I then computed the VAD scores for the sentences in my dataset containing an occurrence of the mentioned aspects. I used weighted average on all sentences collected for each aspect cluster to compute the final VAD scores for the support and against aspect clusters. For consistency reasons, I select key-concepts only when their frequency (in both support and against aspect clusters) $>$ 1,000.

ABSA Pipeline. Figure 3.1 visualizes the proposed pipeline of ABSA applied to polarized context. First, the newspapers articles with the emotions annotated by the users are collected from the Rappler website. Second, a classifier is trained and tested on such dataset (see Section 3.1 for a detailed discussion) to classify the emotions associated to the articles. Then a regression based on each sentence of a polarized news source is addressed to obtain all emotions within each news source. It must be noticed that this

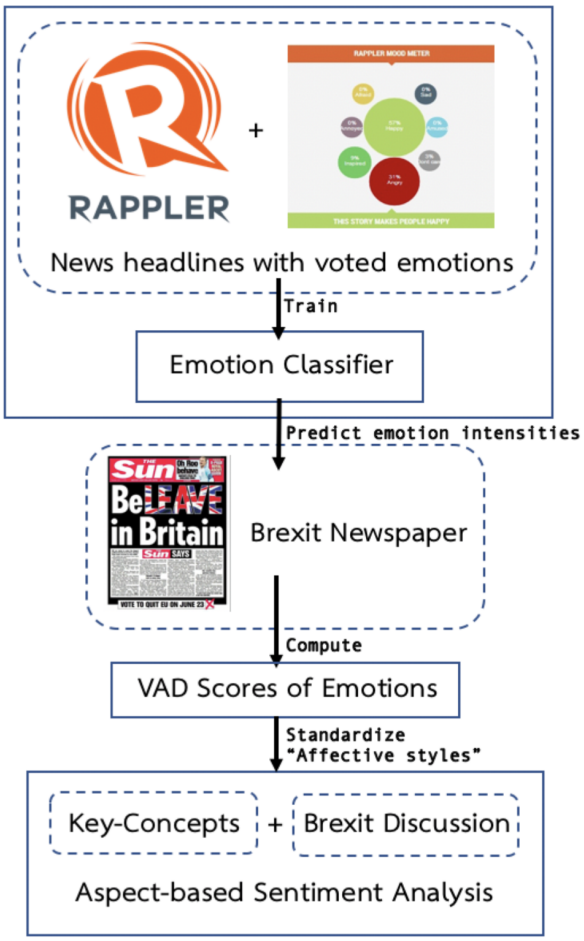


Figure 3.1: The proposed pipeline for ABSA in a polarized context scenario.

pipeline applies the regression to obtain emotions on polarized news source individually. Finally, VAD values are calculated (see Section 3.2 for more details) for each polarized news source. The ABSA model using key-concepts and the Brexit discussion articles are taken into account in the calculation and the visualization of polarization.

Results and discussion. Table 3.2 shows some clear-cut results for key-concepts "Brexit", "Immigration" and "Theresa May". Considering "Brexit" key-concept, I see that the valence of support aspect in pro-Brexit journal is positive, while it is negative in anti-Brexit (as expected). Turning to arousal, I see that the concept "Brexit" provokes a higher emotional activa-

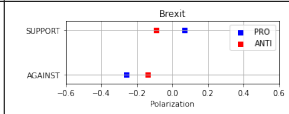
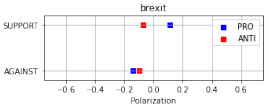
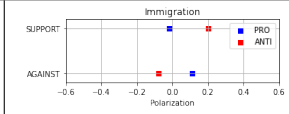
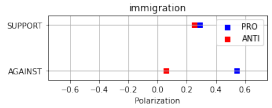
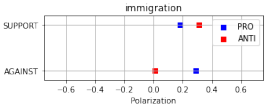
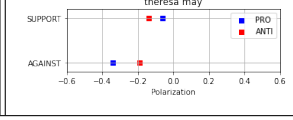
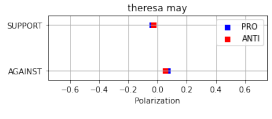
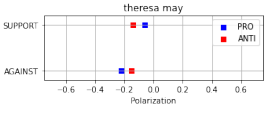
Key-concept	Valence	Arousal	Dominance
Brexit			
Immigration			
Theresa May			

Table 3.2: Polarization on VAD Dimensions using Key-Concepts Approach.

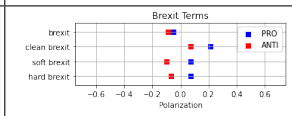
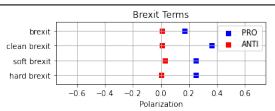
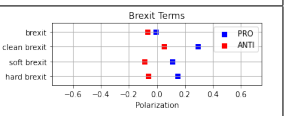
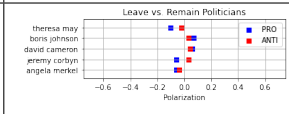
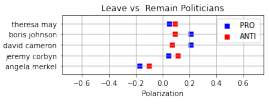
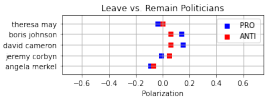
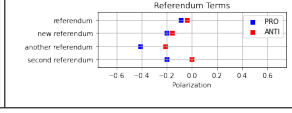
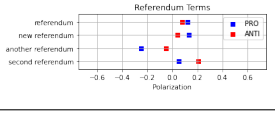
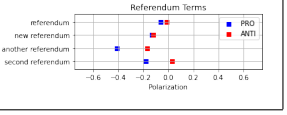
Sub-concept	Valence	Arousal	Dominance
Brexit			
Politician Name			
Referendum			

Table 3.3: Polarization on VAD Dimensions on Sub-concepts via Averaging Topic.

tion in pro-Brexit journal, regardless of the aspect being support or against. "Immigration", instead, is the key-concept with the highest arousal in all the considered cases, both for pro- and for anti- Brexit in the support case, but with opposite valence (positive for anti and neutral/negative for pro). In the against aspect instead (e.g., "stop immigration"), pro-Brexit have an extreme arousal connected to positive sentiment (valence), while anti-Brexit have low arousal with slightly negative sentiment (i.e., a resignation attitude). Interestingly, dominance (the feeling of being in control of the situation) is consistent with the polarized sentiment in both aspects and journals, with

the "resignation attitude" confirmed also in this case for the against aspect of anti-Brexit. Lastly, I consider a politician, "Theresa May", giving that at the time I crawled the articles she was the UK Prime Minister. Results show a convergence toward a neutral score on arousal for both pro- and anti-Brexit (both for support and against key-concepts). For what concerns valence and dominance, they are both in the negative area of the polarization: the anti-Brexit journal has a more negative valence in the cluster of the support aspect (i.e., they are against Theresa May), while the opposite holds for the against aspect cluster, where the pro-Brexit journal has, as expected, more negative valence. These results seem to imply that the pro-Brexit journal has a "lukewarm attitude" toward Theresa May (neutral arousal, almost neutral sentiment in support but negative sentiment when she is criticized).

Finally, in Table 3.3, I analysed some concepts for which I did not have enough occurrences to create the support and against aspect clusters, but are nonetheless important with respect to the journal stance. Referring to sub-concepts of the Brexit scenario, i.e., "clean Brexit", "soft Brexit" and "hard Brexit", I see a clear consistency on all VAD dimensions with the previous analysis. It is also interesting to note that without using the proposed aspects, I obtain a neutral position on the Brexit key-concept alone, and inexact positions on Theresa May as well (compare the two key-concepts in Tables 3.2 and 3.3). This proves that in polarized settings, applying just standard SA to perform ABSA can be misleading for concepts with high occurrences, which is mostly the case in traditional sentiment extraction. Other sub-concepts of Brexit, such as "clean Brexit", "soft Brexit" and "hard Brexit", have highlighted the consistency of VAD trends such that the pro-Brexit journal reveals higher scores on all VAD elements for these sub-concepts, than the anti-Brexit one. The same holds for sub-concepts concerning "politician name" (i.e., pro-Brexit politicians have a positive valence in pro-Brexit journals and anti-Brexit politicians have better valence in anti-Brexit journals). Finally, also for "referendum" anti-Brexit journal shows a consistently higher valence (i.e., preference toward an additional referendum).

3.4 Conclusions

In this chapter, I presented a methodology to extend the task of Aspect-based Sentiment Analysis so that to include affect and emotion representation in polarized settings. In particular, I adopted the three-dimensional model of affect based on Valence, Arousal, and Dominance. I tested the proposed framework on the Brexit scenario, showing how affect dimensions (i.e., VAD) vary on the same aspect when politically polarized stances are presented. The approach proposed in this chapter is able to capture stereotypical aspect-based polarization from newspapers regarding the Brexit scenario using biased key-concepts (e.g., "stop Brexit" vs. "support Brexit").

In addition, the aspect-based sentiment analysis approach presented in this chapter comprises several sentiment elements (i.e., polarity, emotions and VAD) showing that they all play a key role in distinguishing polarization and trends in political news articles. These results highlight that using such sentiment elements can bring up useful guidelines to better characterize such texts which are similar to propagandist ones. Finding out these sentiment elements is proved in this chapter to facilitate the identification of such messages that contain a 'hidden agenda' in public posts.

Chapter 4

Investigating Semantic and Argumentative Features for Supervised Propagandist Message Detection and Classification

One of the mechanisms through which disinformation is spreading online, in particular through social media, is by employing propaganda techniques. This chapter presents and discusses the results obtained on text representation, and classification architectures for propagandist text. These include specific rhetorical and psychological strategies, ranging from leveraging on emotions to exploiting fallacious arguments. In this chapter, the goal is to push forward research on propaganda detection based on text analysis, given the crucial role these methods may play to address this main societal issue. The propaganda detection task is addressed in two steps: first, I performed an investigation of feature representation to characterize propagandist text, and second, I proposed some fine-grained classification architectures to

automatically detect propaganda in text. More precisely, I propose a supervised approach to classify textual snippets both as propaganda messages and according to the precise applied propaganda technique, as well as a detailed linguistic analysis of the features characterising propaganda information in text (e.g., semantic, sentiment and argumentation features). Extensive experiments conducted on two available propagandist resources (i.e., NLP4IF'19 and SemEval'20-Task 11 datasets) show that the proposed approach, leveraging different language models and the investigated linguistic features, achieves promising results on propaganda classification, both at sentence- and at fragment-level. This chapter describes the results published at the International Conference on the Recent Advances in Natural Language Process (RANLP-2021) [77]

Propaganda represents an effective, even though often misleading, communication strategy to promote a cause or a viewpoint, for instance in the political context [29; 49; 51; 54]. Different communication means can be used to disseminate propaganda, i.e., textual documents, images, videos and oral speeches. The ability to effectively identify and manifestly label such kind of misleading and potentially harmful content is of primarily importance to restrain the spread of such information to avoid detrimental consequences for the society.

In this chapter, I tackle this challenging issue pointed out by Da San Martino et al. [23] by proposing a textual propaganda detection model. More precisely, I address the following research questions: *(i) how to automatically identify propaganda in textual documents and further classify them into fine-grained categories?*, and *(ii) what are the linguistic distinctive features of propaganda text snippets?* The contribution of this chapter consists not only in proposing a new effective neural architecture to automatically identify and classify propaganda in text, but I also present a detailed linguistic analysis of the features characterising propaganda messages.

Before defining neural architecture to automatically detect and classify

propagandist text, there is the need to investigate a set of features that tailor such texts in order to employ such influential information in the system. The feature analysis addressed in this chapter is based on psychology (Doob [30]) and computational linguistics studies on propaganda [39; 73; 82]. In this chapter, I propose a fine-grained feature investigation where the simplicity of the message and the argumentation features proposed by Durmus and Cardie [33] are inspected to help boosting the representation of text information.

This chapter focuses on the propaganda detection and classification task, casting it both as a binary and as a multi-class classification task, and I address it both at sentence-level and at fragment-level. I investigate different architectures of recent language models (i.e., BERT, RoBERTa), combining them with a rich set of linguistic features ranging from sentiment and emotion to argumentation features, to rhetorical stylistic ones. The extensive experiments I conducted on two standard benchmarks (i.e., the NLP4IF'19 and SemEval'20-Task 11 datasets) show that the proposed architectures achieve satisfying results, outperforming state-of-the-art systems on most of the propaganda detection and classification subtasks. An error analysis discusses the main sources of misclassification. Furthermore, I analysed how the most relevant features for propaganda detection impact the fine-grained classification of the different techniques employed in propagandist text, revealing the importance of semantic and argumentation features.

The structure of this chapter is the following. Section 4.1 presents the propaganda detection and classification task, and the investigation of the features representation is conducted in Section 4.2. Section 4.3 presents the performance of the proposed features on a baseline sentence classification task. Then, the fine-grained propaganda classification on fragment-level is addressed in Section 4.4. Finally, Section 4.5 discusses the obtained results, and future work directions in propaganda detection.

4.1 The Propaganda Detection and Classification Task

Da San Martino et al. [21] define the Fine-Grained Propaganda Detection task as two sub-tasks, with different granularities: *i) Sentence-Level Classification task (SLC)*, which asks to predict whether a sentence contains at least one propaganda technique, and *ii) Fragment-Level Classification task (FLC)*, which asks to identify both the spans and the type of propaganda technique.

In the following example, *"In a glaring sign of just how stupid and petty things have become in Washington these days, Manchin was invited on Fox News Tuesday morning to discuss how he was one of the only Democrats in the chamber for the State of the Union speech not looking as though Trump killed his grandma."* the span *"stupid and petty"* carries some propagandist bias, and is labeled as "Loaded Language" , *"not looking as though Trump killed his grandma"* is considered as "Exaggeration and Minimisation", and *"killed his grandma"* is "Loaded Language". According to the SLC task, the whole sentence should be classified as a propaganda message given that it contains at least one token (e.g., "*stupid and petty*") considered as such.

As previously introduced, current systems address these tasks relying on word embedding models (e.g., BERT-embedding) and standard features (e.g., PoS, name-entity, n-grams), as representations to feed various RNN architectures [57; 19]. Recently, the language model BERT (Devlin et al. [28]) has been widely utilized to optimize the performances of classification tasks, but there is still room for improvement, in particular when applied to propaganda detection [22; 21]. In this chapter, I experiment with multiple architectures and language models to classify propagandist messages on both sentence and fragment-level. Prior to that, I conduct a detailed investigation of linguistic and argumentation features to capture propaganda strategies.

4.2 Feature Analysis

Propaganda strategies generally involve specific targets to be stimulated by the message. To better study such techniques from a computational point of view, I investigate a set of features that I assume to play a role in propaganda.

4.2.1 Persuasion

Speech style. To analyze the writing style of the messages, I apply the dictionary-based mapping tool "General Inquirer (v. 1.02)" developed by Gilman [40]. It relies on a list of lexicons from 26 domains (e.g., politician speeches, consumer protests) annotated according to 182 rating categories and dimensions (e.g., valence categories and words indicating overstatement and understatement)[1]. I apply such tool on the data and then sum the ratings of each token to obtain a global score for a sentence.

Lexical complexity. Given that pre-trained language models have shown to capture lexical and semantic complexities of words, I rely on BERT (base-uncased) (Devlin et al. [28]) to extract lexical complexity features. I extract a vector of 768 dimensions per each token, then I average w.r.t. all tokens in a sentence, to obtain one vector of 768 dimensions to represent a sentence.

Concreteness. Propaganda messages tend to employ words with concrete meaning, that has more impact in conveying the intention of the message than using abstract words (Eliasberg [36]) I rely on the concreteness lexicon by Brysbaert et al. [16] and sum the standardized score of each token in a sentence to obtain the global score.

Subjectivity. I rely on the subjectivity lexicon from Wilson et al. [82]. I sum up the scores over all tokens in a sentence found in the lexicon as my extracted feature. Each word labeled as "weaksubj" is set to 0.5, and "strongsubj" to 1.

[1] `http://www.wjh.harvard.edu/~inquirer/homecat.htm`

4.2.2 Sentiment

Sentiment labels. I use SentiWordNet 3.0 proposed by Baccianella et al. [6] to obtain word-level sentiment labels (positive, negative, or neutral). I sum the sentiment scores of each word in a sentence, producing a vector with 3 dimensions (i..e, pos, neg, neu) for each sentence.

Emotion labels. I extract 8 emotions (i.e., afraid, amused, angry, annoyed, don't care, happy, inspired, sad) from the DepecheMood++ lexicon (Araque et al. [4]). For each word that evokes emotions in a sentence, I produce the features by summing up each set of emotions evoked by each token, then find the average by emotions. Hence, I produce 8 emotion scores for a sentence.

VAD labels. In the three-dimensional model of affect, valence ranges from unhappiness to happiness and expresses the pleasant or unpleasant feeling about something, arousal expresses the level of affective activation, ranging from sleep to excitement, and dominance reflects the level of control of the emotional state, from submissive to dominant. I use the Warriner lexicon (Warriner et al. [80]) to match each word in a sentence to its VAD standardized word scores and sum up as the features.

Connotation. Propaganda can convey sentiment beyond its original meaning. Connotation lexicon, Feng et al. [38] provide positive, negative and neutral labels of each word. I count the frequencies of the three labels evoked in each sentence.

Politeness. Politeness evokes sentiment in readers. I use a lexicon of positive and negative words from Danescu-Niculescu-Mizil et al. [26], then I count the frequencies of both positive and negative words found in each sentence.

4.2.3 Message Simplicity

To keep the message simple and picturable is one of main purposes of propaganda. I list the features to extract the simplicity of the message dimension.

Exaggeration. I use imageability lexicon from Tsvetkov et al. [74] based on picturable vocabulary which mentally leads to an exaggerating state of

mind. I consider the scores of abstraction and concreteness at each word token. I then sum up the scores for all the labels found in a sentence.

Length. "The less words used, the better to understand" can be a concept to easily interpret the propagandist message. I apply two strategies: *i)* I count the average char-length, actual char-length, word length, punctuation frequency, capital-case frequency per sentence (Ferreira Cruz et al. [39]); *ii)* I apply length encoding at character-level, plus one additional dimension for non-alphabetical char count.

Pronouns. Loaded language, name calling and labelling are the most used techniques in propaganda text (Da San Martino et al. [21]), and they all make use of pronouns. I create a lexicon of 123 pronouns in English[2] and perform one-hot encoding of common used pronouns in English.

4.2.4 Argumentation

Argumentation may play an important role in propaganda [33]. To extract argumentative features representing the data, I train a supervised classifier for the task of argumentative sentence classification on the persuasive essays dataset (Stab and Gurevych [69]). First, I cast it as a binary classification task, merging premises, claims and major claims into the *argumentative* label, as opposed to the *non-argumentative* label. Then, for the argumentation component task, I rearrange the data to binary labels where the major claims and claims labels are merged, and they are opposed to premises. To address these tasks, I build and fine-tune a BERT classifier. I use a learning rate of 1e-5 with 80/20 split of the dataset. I run a classifier 3 times at different random states. The results for the argumentative sentence classification are (macro-average) F1 0.84, precision 0.86, recall 0.82, while for the component classification they are F1 0.77, precision 0.80, recall 0.75.

To extract argumentative features from the annotations provided by the classifiers, I use BERT-based features. After fine-tuning, I freeze the hidden states of these fine-tuned BERT models. To extract the argumentative and

[2]https://www.englishclub.com/vocabulary/pronouns-type.php, https://www.thefreedictionary.com/List-of-pronouns.htm

Persuasion	Sentiment	Message simplicity	Argumentation	Logistic Regression			BERT + Featured LR		
				F1	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
✓				0.68	0.69	0.67	0.70	0.71	0.70
	✓			0.62	0.69	0.57	0.71	0.72	0.69
		✓		0.63	0.66	0.62	0.70	0.72	0.68
			✓	0.67	0.68	0.67	0.71	0.72	0.70
✓	✓			0.68	0.70	0.67	0.71	0.71	0.69
✓		✓		0.69	0.71	0.68	0.70	0.71	0.69
✓			✓	0.69	0.70	0.68	0.70	0.71	0.70
	✓	✓		0.66	0.68	0.65	0.71	0.73	0.70
	✓		✓	0.70	0.71	0.69	0.71	0.72	0.70
		✓	✓	0.66	0.68	0.65	0.70	0.72	0.69
✓		✓	✓	0.69	0.70	0.68	0.71	0.72	0.70
✓	✓		✓	0.68	0.69	0.68	0.70	0.71	0.69
	✓	✓	✓	0.70	0.72	0.69	0.70	0.71	0.69
✓	✓	✓		0.71	0.72	0.69	**0.72**	**0.74**	0.70
✓	✓	✓	✓	0.70	0.71	0.69	0.71	0.72	0.69

Table 4.1: Ablation test on binary classification setting.

components features from each classifier, I take the [CLS] token of each sentence from the fine-tuned BERT model.

4.2.5 Ablation Study

To investigate the impact of the proposed features (Section 4.2) for propaganda detection, I perform ablation tests by testing a supervised classifier relying on BERT + logistic regression. To the purpose, I use the NLP4IF'19 training and test sets (Da San Martino et al. [21]).

Table 4.1 reports on the performances obtained while integrating groups of features to the proposed model. A logistic regression model is used as a baseline. Best results are obtained when adding all the proposed features, but the argumentation ones. Argumentation features alone perform almost identical as semantic features, therefore - unexpectedly - no added value can be demonstrated.

4.3 Sentence-level Classification

In the following, I describe the experiments I carried out to address the propaganda detection task at sentence level, investigating different architectures and leveraging both recent language models and the features that proved to play a role in propaganda messages. For the evaluation, I used the two available datasets for propaganda detection: *i)* the NLP4IF'19 data set (Da San Martino et al. [21]) (293 articles for training and 101 for testing); and *ii)* the data from SemEval'20 T11 (Da San Martino et al. [22]) (371 articles for training and 75 in the development set).

Model	NLP4IF'19 Test Set			SemEval'20-T11 Dev. Set		
	F1	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
BERT ***Baseline***	0.52	0.53	0.50	0.48	0.48	0.48
SOTA						
Fine-tuned BERT	0.58	0.63	0.53	0.61	0.63	0.60
Fine-tuned T5	0.64	0.64	0.65	0.66	0.65	0.66
Linear-Neuron Attention BERT	0.63	0.60	0.67	0.66	0.69	0.63
Multi-granularity BERT	0.61	0.60	0.62	0.65	0.68	0.63
Proposed Architecture w/ Semantic Features						
Multi-granularity + Featured BERT	0.63	0.65	0.61	0.67	0.71	0.64
BERT + Featured BiLSTM	0.65	**0.80**	0.55	0.65	**0.75**	0.58
BERT + Featured Logistic Regression	**0.72**	0.74	**0.70**	**0.68**	0.71	0.66
Proposed Architecture w/ Semantic Features + Argumentation Features						
BERT + Featured Logistic Regression	0.71	0.72	0.69	**0.68**	0.70	**0.67**

Table 4.2: Results on the Sentence-level classification (SLC) task (binary task).

4.3.1 Prediction Models

In the following, I first describe the baseline and the SOTA models I tested in the experiments, and then I present the three architectures I propose

(underlined) integrating the propagandist features previously investigated (Section 4.2).

BERT. The baseline model relies on a pre-trained bidirectional transformer language model to encode context specific sentence tokens (Devlin et al. [28]) (no fine tuning, default hyperparameters).

Fine-tuned BERT. I fine-tune the BERT model with a learning rate of 5e-5, and AdamW optimizer. I set the gradients to zero at every training batch. Then I use softmax activation to gate the output with the threshold of 0.5.

Fine-tuned T5. To fine-tune the text-to-text transformer proposed by Raffel et al. [60], I use T5 For Conditional Generation approach (equally to question-answering task) where the input is a sentence (as a question), and the output is an answer (as a label). I use a learning rate of 3e-4, with max sequence length of 512.

Linear-Neuron Attention BERT. I replicate the winning approach of the NLP4IF'19 shared-task (Mapes et al. [55]). It relies on BERT architecture with some modifications of hyperparameters (sentence length of 50 tokens, a learning rate of 1e-5, along with 12 attention heads and 12 transformer blocks). It uses the linear neuron without an activation function, and a threshold of 0.3 for the final prediction.

Multi-granularity BERT. The model from Da San Martino et al. [21] relies on BERT transformer with multi-granularity network on top that has multi-classifiers for different granularity levels of text (e.g., document, paragraph, sentence, word, subword, and character-level). I replicate this model with BertAdam optimizer and ReLU activation function.

Multi-granularity + Featured BERT. I integrate the proposed features (Section 4.2) into the model proposed by Da San Martino et al. [21], taking only the last layer of sentence-level granularity. I feed the proposed features to a BERT classifier to obtain logits which then aggregate with the last layer of sentence-level granularity to produce predictions.

BERT + Featured BiLSTM. I build a pre-trained BERT transformer architecture, and Bi-directional Long Short-Term Memory (BiLSTM) architecture on top of the BERT model to handle the transformer architecture with

my propaganda features. Firstly, the BERT model is used with learning rate of 0.001, with AdamW optimizer. I use the output of BERT that represents the [CLS] token of each sentence to combine with propaganda features as the input to the second model, the BiLSTM. For the BiLSTM model, after I feed the inputs of both [CLS] tokens combined with propaganda features, I train the BiLSTM model with hidden size of 256. The BiLSTM hidden states consist of the last hidden states, and the last cell state for the BiLSTM layers. I then apply relu gate function, with a linear dense, then use a dropout function of 0.1. At the last layer, I use another linear dense layer to output final logits, then I apply a sigmoid activation function as final outputs.

BERT + Featured Logistic Regression. I use pre-trained BERT transformer architecture to output [CLS] token, then use this output to stack with another prediction model, i.e., logistic regression. I build a linear classifier and feed it with propaganda features as a dense layer. I then combine these logits with [CLS] tokens as the input to logistic regression on top of BERT.

4.3.2 Results and Error Analysis

False Positive	**False Negative**
People who hate freedom will get unfettered access to the minds of 2 billion people.	The American people have a right to know, and those that engaged in this type of behavior do not have a right to hide.
You are a slave to white America.	Hitler was a very great man.
So proud to support Tommy Robinson and free speech in London today.	87 is the average Indonesian IQ , and note that average includes the higher average Chinese, so the locals really are a dull lot.
Shame on those who are supposed to uphold law and justice!	Earlier, I blogged that the police had released the name of the suspect in the murder of two white Westerville, OH cops (Quentin Lamar Smith) but no picture had appeared.

Table 4.3: Examples of misclassified sentences by the BERT + featured logistic regression model (NLP4IF'19)

Table 4.2 reports on the results obtained for the SLC task (*propaganda*

vs *no propaganda*). I run each experiment 5 times and report the macro-average of all metrics. The proposed models achieve the highest F1-score of 0.72 using BERT + Featured Logistic Regression model (persuasion, sentiment, and message simplicity features), and the highest precision-score 0.80 using BERT + Featured BiLSTM model on NLP4IF'19 dataset, outperforming the state-of-the-art models. For SemEval'20-T11, I do not have the scores from the challenge (the binary task was not proposed), but I compare the obtained results with the replicated architectures of SOTA models. The proposed architecture obtained the best F1-score using BERT+Featured Logistic Regression. Using semantic features alone perform slightly better than combining them with argumentation features.

Table 4.3 reports on some misclassified examples of the best model on NLP4IF'19 dataset. Some short sentences containing strong intention keywords (e.g., "hate", "slave") have been missclassified as false positives. As for false negatives, the underlined fragments are labeled propaganda in gold standard, but have not been recognized as such by the classifier (mainly informative statements).

4.4 Fragment-level Classification

In this section, I address the task of fragment-level classification, meaning that both the spans and the type of propaganda technique should be identified in the sentences. Again, to test the proposed methods, I use both NLP4IF'19 and SemEval'20 T11 datasets. However, in the two challenges, the FLC task was evaluated according to different strategies, explained in the following.

4.4.1 Task 1: FLC on NLP4IF'19 Dataset

In the NLP4IF'19 dataset, 18 propaganda techniques are annotated. Prediction is expected to be at token-level. Multiple tokens can belong to the same span, and annotated with one propaganda type. Tokens that do not carry any propaganda bias are annotated as "no propaganda". To perform

tokenization I run the tokenizer provided with the pretrained model of each transformer[3].

Prediction Models

Fine-tuned BERT (baseline). Pretrained *bert-base-uncased* model and BERT architecture (Devlin et al. [28]) with default hyperparameters. An implementation is based on huggingface transformers. Settings: learning rate of 5e-5, padded length of 128, and batch_size of 16. I use CrossEntropyLoss as a loss function, and softmax activation function to gate output neurons.
Fine-tuned RoBERTa (baseline). I use *roberta-base* model with the same hyperparameters of loss and activation functions as the fine-tuned BERT model mentioned above.
State-of-the-art Model. The winning team applied BERT architecture for token classification (Yoosuf and Yang [84]) on 20 labels (i.e., 18 propaganda classes, plus "background" as non propaganda, and "auxiliary" for fractions of previous tokens). They use a BERT language model, then apply softmax function, followed by a linear multi-label classification layer to output their predictions.
Transformer + CRF. I use a pre-trained model *base-uncased* with a learning rate of 3e-5 for BERT transformer, and a pre-trained model *roberta-base* with a learning rate of 2e-5 for RoBERTa transformers (hyperparameters: dropout of 0.1 with the max_length of 128, batch_size of 16 with AdamW optimizer and CrossEntropy loss function). I use CRF layer as the final layer.

Results and Error Analysis

Table 4.4 reports on the obtained performances. Evaluation is reported as the average of micro-F1 scores of 5 run-times (I use the evaluation scripts provided by Da San Martino et al. [21]).

The proposed architecture based on transformers with CRF output layer at different learning gradients (epochs) outperforms SOTA model on several propaganda techniques at different learning gradient ranging from 5 to 15

[3]huggingface.co/transformers/

	NLP4IF'19											
	Average	Appeal-Fear	Black-White	Casual-Over.	Doubt	Exag.-Min.	Flag-Waving	Loaded-L.	Namecalling	Reductio-Hit.	Repetition	Slogans
Baseline												
Fine-tuned BERT	0.03	0.09	0.04	0.03	0.07	0.07	0.17	0.08	0.07	0.02	0.01	0.04
Fine-tuned RoBERTa	0.02	0.06	0.02	0.01	0.05	0.07	0.10	0.09	0.07	0.01	0.01	0.02
SOTA (from NLP4IF'19) [84]	0.24	0.21	0.09	.0	0.17	0.16	0.44	0.33	0.40	.0	0.01	0.13
Proposed Architecture												
Fine-tuned BERT + CRF (5 epochs)	0.13	0.27	.0	0.04	0.08	0.20	0.59	0.26	0.28	**0.08**	0.01	0.10
Fine-tuned BERT + CRF (15 epochs)	0.11	0.25	0.02	0.04	0.07	0.28	**0.61**	0.25	0.22	0.04	**0.04**	0.13
Fine-tuned RoBERTa + CRF (5 epochs)	0.16	0.32	.0	**0.09**	0.11	0.35	0.37	**0.42**	0.37	.0	.0	0.06
Fine-tuned RoBERTa + CRF (7 epochs)	0.14	**0.40**	**0.23**	0.08	0.13	**0.37**	0.46	0.37	0.31	0.05	.0	0.17
Fine-tuned RoBERTa + CRF (10 epochs)	0.15	0.30	0.19	**0.09**	0.13	0.31	0.53	0.35	0.29	.0	0.01	0.25
Fine-tuned RoBERTa + CRF (12 epochs)	0.15	0.31	0.17	0.05	**0.19**	0.31	0.47	0.33	0.32	.0	0.03	0.14
Fine-tuned RoBERTa + CRF (15 epochs)	0.16	0.35	0.16	0.03	0.16	0.35	0.49	0.33	0.27	.0	0.01	0.24
Fine-tuned RoBERTa + CRF (5 epochs 3x-Oversampled)	0.15	0.34	0.14	0.07	0.13	0.30	0.52	0.34	0.27	0.05	0.02	**0.33**

Table 4.4: Experimental results on fragment-level classification on NLP4IF'19 test set. All scores are reported in micro-F1 (as in the original challenge). Scores in bold are the ones outperforming SOTA model.

epochs. I also tested other architectures such as Transformer+CRF with less learning gradients (3 epochs), Transformer architecture with semantic and/or argumentation features + CRF layer by adding extracted features from sentence-level (Section 4.2) to each token of its sentence to a linear layer before a loss function, with no major improvements.

In Table 4.4, I compare the performances of the proposed models w.r.t. the SOTA (Yoosuf and Yang [84]), on the most frequent classes. Table 4.5 reports examples of misclassification related to that technique. I observe that the proposed model does not capture well the articles (i.e., it, as, an, the), but rather focuses on capturing intentional word tokens (i.e., white, unbelievably, rude, wonderful, treasonous). As for future work to improve results on this specific category, I will investigate the work of Habernal et al. [44] according to which a dedicated strategy is needed.

4.4.2 Task 2: FLC on SemEval'20 T11 Dataset

In SemEval'20 T11 dataset, 14 propaganda techniques are annotated. I focus here on the task called Technique-Classification task (TC). I cast it as

Technique: NameCalling_Labeling
(a) We look forward to continuing to defend the White House's lawful actions.
(b) This is a man who follows a demonic religion, Islam, and supports textcol-orredcop killers.
(c) It was not widely recognized as an anti-Jewish organization during its early years (its early literature, though, focused on "the white man" as "the white devil").
(d) The president described Acosta as "unbelievably rude to [White House Press Secretary] Sarah Huckabee, who's a wonderful woman," and said his administration is drawing up "rules and regulations" for White House reporters.
(e) She is like Derrick Kahala Watson, a treasonous kritarch, unwilling to submit her hunger for power and ideology to the Constitution [...].'

Table 4.5: Misclassified NameCalling_Labeling. False Negative (in red), False Positive (in blue), the correctly classified propaganda spans (underlined).

a sentence-span classification problem, where I combine logits of tokenized elements from the sentence and the span, to learn the prediction. Moreover, I add the semantic and argumentation features to enhance the performance.

As pre-processing, both the tokenized sentence and the span are used to feed the transformer (Huggingface tokenizers) as follows: *i)* I input a sentence to the tokenizer where max_length is set to 128 with padding; *ii)* I input the span provided by the propaganda span-template published by the workshop, and I set max_length value of 20 with padding. If a sentence does not contain propaganda spans, it is labeled as a "none-propaganda".

Prediction Models

Baseline. For all the tested architectures (BERT and RoBERTa), I use the same type of transformer model to produce logits (L) regarding the sentence-level and span-level individually. For BERT model, I use pre-trained model *bert-base-uncased*, learning rate of 5e-5, and α of 0.1. For RoBERTa, I take *roberta-base* pre-trained model with learning rate of 2e-5 with α of 0.5. All transformer models apply Adam optimizer, dropout 0.1, and CrossEntropy

	SemEval'20 T11														
	Average	Appeal_to_Authority	Appeal_to_fear-prejudice	Bandwagon,Reductio_ad_hit.	Black-White-Fallacy	Casual-Oversimplification	Doubt	Exaggeration,Minimisation	Flag-Waving	Loaded_Language	Name_Calling,Labeling	Repetition	Slogans	Thought-terminating_Cliches	Whatab.,Straw_Men,Red_Her.
SOTA (from SemEval'20 T11) [47]	0.64	0.48	0.47	0.08	0.51	0.23	0.56	0.37	0.70	0.78	0.76	0.59	0.59	0.39	0.28
Proposed Architecture + Proposed argumentation features															
Fine-tuned RoBERTa (3 epochs)	0.53	0.08	0.34	0.14	0.17	0.06	0.52	0.32	0.61	0.72	0.68	0.22	0.12	**0.42**	.0
Fine-tuned RoBERTa (5 epochs)	0.53	0.14	0.34	0.17	0.26	0.09	0.46	0.35	0.60	0.73	0.72	0.17	0.36	0.30	0.18
Fine-tuned RoBERTa (10 epochs)	0.51	0.18	0.33	0.13	0.37	0.22	0.37	0.33	0.58	0.73	0.68	0.17	0.34	0.17	0.23
Fine-tuned RoBERTa (15 epochs)	0.51	0.14	0.29	0.12	0.31	0.14	0.42	0.35	0.55	0.73	0.69	0.13	0.35	0.25	0.21
Proposed Architecture + All proposed features															
Fine-tuned RoBERTa (3 epochs)	0.54	0.16	0.38	**0.20**	0.29	0.18	0.50	0.33	0.60	0.72	0.65	0.23	0.29	0.32	0.09
Fine-tuned RoBERTa (5 epochs)	0.52	0.09	0.35	0.13	0.31	0.21	0.43	0.34	0.61	0.74	0.70	0.21	0.23	0.33	0.12
Fine-tuned RoBERTa (10 epochs)	0.51	0.09	0.31	0.17	0.37	0.28	0.36	0.35	0.54	0.73	0.70	0.19	0.38	0.14	0.19
Fine-tuned RoBERTa (15 epochs)	0.51	0.15	0.32	0.07	0.40	**0.29**	0.37	0.31	0.54	0.75	0.66	0.18	0.43	0.14	0.12

Table 4.6: Results on span classification on SemEval'20 T11 test set (micro-F1).

as a loss function per sentence ($loss_{\text{sentence}}$) and span ($loss_{\text{span}}$).

I arrange these alignment of L to calculate the average loss as joint loss ($loss_{\text{joint\_loss}}$) from each *loss* element. Here I introduce a $loss_{\text{joint\_loss}}$ function before back-propagation:

$loss_{\text{joint\_loss}} = \alpha \times \frac{(loss_{\text{sentence}} + loss_{\text{span}})}{N_{\text{loss}}}$ where N_{loss} stands for a number of *loss* elements that are taken into the model.

State-of-the-art Model. Jurkiewicz et al. [47], authors of the winning team apply RoBERTa (*roberta-large*) with pre-trained model. The training set is increased with silver annotation based on gold annotation, and then another RoBERTa model is stacked on top to output the predictions.

Proposed Architecture. I propose another set of elements to feed the transformer by introducing the semantic and argumentation features into BiLSTM layer to produce L of proposed features, then I apply CrossEntropy as a loss function of the BiLSTM as $loss_{\text{proposed\_features}}$ then perform an addition with other *loss* in the $loss_{\text{joint\_loss}}$ function as follows: $loss_{\text{joint\_loss}} = \alpha \times \frac{(loss_{\text{sentence}} + loss_{\text{span}} + loss_{\text{proposed\_features}})}{N_{\text{loss}}}$ Hyper-parameters: 256 hidden_size, 1 hidden_layer, drop_out of 0.1 with ReLU function at the last layer before the joint loss function.

Results and Error Analysis

Technique: Repetition	False Negative
No Extra Feature	
(1) Is it simply a coincidence that this, researcher, prior to coming to Oxford University, worked for Tell Mama, that factory for the production of bogus claims about Islamophobia?	Name_Calling,Labeling
(2) And why are Muslims allowed to cover their faces with a black sack while the rest of us are subject to strict security?	Doubt
(3) The boy told agents that Ibn Wahhaj trained him and another of Leveille, teenage sons in firearms and military techniques, including rapid reloads and hand-to-hand combat, and told them jihad meant killing non-believers on behalf of Allah, according to the affidavit filed in U.S. District Court in New Mexico.	Appeal_to_fear-prejudice
Semantic Feature	
(1) Why is it so easy for judges to make rulings that allow known terrorists and jihadists to stay in our country?	Doubt
(2) Not even the richest people in the country with the best lawyers would receive this type of treatment if they just admitted to killing an innocent person.	Name_Calling,Labeling
(3) Not even one, he wrote on Twitter.	Exaggeration,Minimisation
Argumentation Feature	
(1) Tommy Robinson is in prison today because he violated a court order demanding that he not film videos outside the trials of Muslim rape gangs.	Name_Calling,Labeling
(2) Even today, there is little value in insuring the survival of our nation if our traditions do not survive with it.	Doubt
(3) In the coming days, we will be filing a major federal lawsuit against the state of Georgia for the gross mismanagement of this election and to protect future elections from unconstitutional actions.	Exaggeration,Minimisation
All Features	
(1) When she arrived at Jean's door, Guyger entered a unique door key with an electronic chip into the keyhole, the affidavit says.	Bandwagon,Reductio_ad_hitlerum
(2) She told the 911 operator as well as responding officers that she thought she was at her apartment when she shot Jean, according to the affidavit.	Doubt

Table 4.7: Misclassified Repetition spans (in red).

As mentioned before, the gold labels of the test set of SemEval'20 T11 are not available, but it is possible to submit a system run to the challenge website and to obtain the evaluation score. The evaluation system only accepts the exact list of span-templates of the test set (partial overlapping spans or missing spans are not accepted). Table 6.1 reports on the obtained results (through such evaluation system) on 5 runs as micro-F1. Scores in bold are the ones for which significant improvement can be observed w.r.t.

SOTA model. RoBERTa with argumentation features can outperform results on "Thought-terminating_Cliches". Moreover, by using all semantic and argumentation features together, I can obtain some improvements over "Bandwagon,Reductio_ad_hitlerum" and "Casual-Oversimplification". Table 4.7 shows some examples of missclassified instances. In general, I noticed that using different training epochs help detecting different propaganda techniques. In particular, it is observed that some techniques tend to be learnt best at low training epochs (i.e., "Bandwagon,Reductio_ad_hitlerum", "Thought-terminating_Cliches"), some at high training epochs (i.e., "Casual-Oversimplification").

4.5 Concluding Remarks

In this chapter, I proposed a new neural architecture combined with state-of-the-art language models and a rich set of linguistic features for the detection of propaganda messages in text, and their further classification along with standard propaganda techniques. Despite the boost in accuracy I achieved on two standard benchmarks for propaganda detection and classification ($\sim$ 10% of F1 scores on sentence-level classification and on specific propaganda techniques on fragment-level classification), this task remains challenging, in particular regarding the fine-grained classification of the different propaganda classes.

Moreover, apart from the higher accuracy obtained with the proposed neural architecture, the propagandist text features are thoroughly investigated and distinctly identified, to highlight the characteristics of propaganda in texts based on linguistic, syntactic, and argumentation attributes. The neural architecture I proposed in this chapter corroborates and demonstrates that using the SOTA of neural architectures alone, without employing any feature peculiar to this kind of linguistic phenomenon (i.e., propaganda), leaves small room for improvement in the propaganda detection and classification tasks. As can be observed from the obtained results discussed in this Chapter, some propaganda techniques have a high F1-score for high training

epochs but it also causes an overfitting.

Starting from the achieved results, the next objective points out onto large scale online text detection and classification to detect, filter, or rank online content with respect to propaganda factors to overcome such disinformation element.

Chapter 5

Propaganda in Political Debates

This chapter describes the annotation process of propaganda techniques as a new annotation layer of an existing data set of political debates. Propaganda plays a central role in political discourse and debates, and the ability of detecting it automatically would be a relevant achievement for the NLP research community. To this purpose, starting from the USElecDeb60To16 data set for argument mining [45], I proposed some annotation guidelines for annotating propaganda in political debates with the goal to add a new annotation layer to this dataset, i.e., the propaganda technique. All annotation and reconciliation phases concerning this new annotation layer are explained in this chapter, as well as the satisfactory inner-annotator agreement we obtained. This chapter concludes with the presentation of the data statistics of the new layer annotated on the data set of political debates. The results presented in this chapter are under submission.

Propaganda in political debates [31; 53] can be produced inadvertently or with the explicit aim of deceiving others using persuasion techniques that are commonly used to influence others and cause errors in reasoning, ridiculing, and being sarcastic. Commonly, propaganda in political debates is used to

emphasize different purposes of the 'hidden agenda', and it is usually found where criticism, illusions, stereotypes, superstition, rationalization, fallacious arguments, or poor judgment are shown [62; 79].

Most of the propaganda techniques I discussed in the previous chapters relate to news, with a strong focus on political issues. This chapter pushes forward this focus and concentrates on a different text genre, namely political debates. More precisely, the goal of this chapter is to present and discuss the whole process of building a new annotated data set of political debates for the propaganda detection and classification tasks.

One of the most representative examples of political debates are the United States presidential debates. In the U.S. presidential election campaign, it is common for multiple candidates to enter a dialogue. The issues discussed within the debate are usually the foremost disputable subjects of the time. These debates targets primarily undecided voters, i.e., people who tend to not believe in any political ideology or party. The formats of the debates are varied, typically guided by the questions asked by journalists or moderators, or, in alternative, by the audience. Debates are broadcast live via television, radio, and on the Internet. On the one hand, I decided to focus on these debates as they contain numerous and representative instances of different propaganda techniques. On the other hand, these debates open new challenges for the propaganda detection and classification tasks as the identification of these instances is more complex than in news articles, due to several reasons, e.g., the dialogical structure of the debate where two candidate are opposing their ideas.

In Section 5.1 I present the annotation guidelines, detailing the propaganda techniques I selected, and the issue of identifying the propagandist text snippet boundaries in the annotation task. Section 5.2 discusses the definition of the different propaganda techniques I selected, through the help of examples from the data set, and exceptions to the annotation guidelines. Section 5.3 explains the annotation and reconciliation phases we addressed as well as the platform used for the annotation. Section 5.4 presents the evaluation procedure conducted during the annotation. Section 5.5 ultimately reports the statistics of the annotated data set of political debates. The

conclusion of the data annotation task and the discussion of the future directions toward propaganda detection and classification in political debates are detailed in Section 5.6.

5.1 Annotation Guidelines

In creating this new data set of debates annotated with propaganda techniques, I started by defining the annotation guidelines specifying the propaganda classes and their definition, the constraints to annotate them in text (e.g., length of the text snippet), and the exceptions, if any. These guidelines provide several examples and the justification explaining why the annotated text snippet fits in that specific category of propaganda. The propaganda techniques I selected in this section have motives based on the concept, notion and intention in targeting the message specific to the context of political debates from Dowden [31] where 230 names of the most common propaganda are listed. After a careful reading of the political debates contained in the USElecDeb60To16 data set [45], I ended up with 6 propaganda techniques to be annotated, namely *Appeal to authority*, *Ad hominem*, *Appeal to emotion*, *False cause*, *Slogans*, and *Slippery slope*. However according to Dowden [31] concerning political debates, there are techniques that partially share the purpose and intention of delivering the message. Some of these techniques can be detailed in sub-categories, ending up with 14 types of fine-grained propaganda techniques (e.g., *Ad hominem*, *Circumstantial Ad hominem*, *Name-Calling Labeling*, *Tu quoque*, *Appeal to popular opinion*, *False Authority*, *Without Evidence*, *Appeal to fear*, *Appeal to pity*, *Flag waving*, *Loaded Language*, *False cause*, *Slippery slope*, *Slogan*). The definition of each propaganda (sub-)category is given within these guidelines.

The annotated data set is a collection of political debates prior to presidential elections in the United States from 1960 to 2016. All the data are transcribed from speech into text-based format and are publicly available via the Commission on Presidential Debates of the United States (CPD)'s web-

site[1]. Each debate is divided into several sections in which the candidates are debating on the same topic. The USElecDeb60To16 dataset [45] comes with annotated labels of argumentation where argument components (evidence, claim) and relations (support, attack) are marked. To the best of my knowledge, it is the biggest available dataset of political debates annotated with argumentation elements.

5.1.1 Standard Rules of Annotation Boundary

In this annotation, the standard rules are set regarding the boundary of spans. There are two annotation levels in this annotation task.

- Sentence-level: if the annotating span wraps the whole sentence, annotator must include all punctuation marks in the annotation.

- Span-level: if multiple words are considered to be annotated as a span (which is a partial words of the whole sentence) where the span is self-explained their meaning and intention toward the propaganda technique, annotator must include only punctuation marks that are aligned within such span. Annotators do not annotate punctuation marks outside the span.

5.2 The Propaganda Techniques

In this section, I detail each propaganda technique we consider in the annotation task, through examples and exceptions. In each example, the underlined text is the text snippet which was annotated with a certain propaganda technique.

[1]debates.org

5.2.1 Appeal to authority (Ad Verecundiam)

In this category, politicians use opinion of experts as their support. Three types of propaganda technique are defined.[2]

Appeal to authority without evidence

Stating a statement is true just because a valid authority said it is true without providing the supporting statement why they said so.

Examples:

- And judgment is what we look for in the president of the United States of America. I'm proud that important military figures who are supporting me in this race: former Chairman of the Joint Chiefs of Staff John Shalikashvili; just yesterday, General Eisenhower's son, General John Eisenhower, endorsed me; General Admiral William Crown; General Tony McBeak, who ran the Air Force war so effectively for his father– all believe I would make a stronger commander in chief [3]

  Justification: The candidate's statement expresses as why he is a good commander in chief is solely based on the approval of the authorities he is mentioning and no other justification of why their judgement is correct is given thus this is considered as appeal to authority.

- if we suffer defeat in Iraq, which General Petraeus predicts we will, if we adopted Senator Obama's set date for withdrawal, then that will have a calamitous effect in Afghanistan and American national security interests in the region[4]

  Justification: The candidate is basing his statement that a defeat in Iraq is inevitable just because an authority said so without providing any other verification, thus he is committing a propaganda of appeal to authority.

[2]Boundary Note: The boundary of this annotation is usually within one or multiple sentences which contains statement(s) the candidate referring to another authority's saying or actions in order to justify their (the candidate) statements.

[3]Bush-Kerry debate , September 30th 2004

[4]Mccain-Obama debate, September 26th 2008

Appeal to false authority

When a false authority's opinion is used as a support to his statement which is not that authority's field of expertise:

Examples:

- Look at her website, she is telling us how to fight ISIS on her website , I don't think general Douglas MacArthur[5] would like that too much. [6]
  Justification: The candidate is trying to mock the opponent's war tactics by mentioning a relevant authority in the field who has been dead for a long time. Thus he is committing a propaganda of appeal to authority.

- But in the case of missile defense, Senator Obama said it had to be, quote,"proven". That wasn't proven when Ronald Reagan said we would do SDI, which is missile defense. And it was major – a major factor in bringing about the end of the Cold War. We seem to come full circle again.[7]
  Justification: The candidate is trying compare how Ronald Reagan handled the missile crisis in his time with the current time. Which is appealing to an authority on the relevant field but not related to the current situation. Thus the candidate is committing propaganda of appeal to authority.

In the first two subcategories of appeal to authority namely false authority and appeal to authority without evidence, the propaganda needs to explicitly say the name of the authority and mention what they have said, supported or opposed which the candidate uses to justify as their support to the statement.

[5] A military general who played an important role in the US army during World War II, he died in 1964

[6] Clinton-Trump debate, September 16th 2016

[7] Mccain-Obama debate, September 26th 2008

Appeal to popular opinion (ad populum)

This propaganda technique covers instances of attempted reinforcement of political statement by referring to the fact that something is very popular, or the will of the people.

In annotating the appeal to popular opinion annotator should look for key terms such as :"people's opinion ", "people's vote" , "majority of people" and terms which imply the same meaning.

Example:

- He can make any excuse he wants, but the facts are that we're reducing the number of uninsured percentage of our population. And as the percentage of the population is increasing nationally, somehow the allegation that we don't care and we're going to give money for this interest or that interest and not for children in the State of Texas is totally absurd. Let me just tell you who the jury is. The people of Texas. There's only been one governor ever elected to back-to-back four-year terms, and that was me[8]

  Justification: The candidate is using his selection as governor to appeal to popular opinion. He is saying: since the majority of people chose me as governor for a four year term, the popular opinion is that I am a good governer thus I am a good governer (who conducts correct insurance laws). Thus the whole statement is annotated as appeal to authority.

Exceptions:

- A statement is not appeal to authority if the candidate explains the justification of the authority of that statement. (Why the authority thinks that statement is true)

- A statement containing the name of an authority is not considered propagandist if the candidate does not mention their name to use as a reason why his statement is justified.

[8]Bush-Gore debate, October 11th 2000:

- In this category of propaganda (for the first two subcategories) the authority should have been named. In the case of popular opinion propaganda the candidate explicitly refer to the majority of people as the authority who accepts them or confirms their actions.

Examples of exception:

- The following example is an exception to appeal to authority because firstly the authority is relevant to the subject and secondly the candidate is providing statements why the authority is correct in their judgement. Thus no propaganda is committed by mentioning the authority.
  Example from Bush-Gore debate, October 11 2000: ... I didn't think he necessarily made the right decision to take land troops off the table right before we committed ourselves offensively, but nevertheless, it worked. The administration deserves credit for having made it work. It is important for NATO to have it work. It's important for NATO to be strong and confident and to help keep the peace in Europe. And one of the reasons I felt so strongly that the United States needed to participate was because of our relations with NATO, and NATO is going to be an important part of keeping the peace in the future...

- In the following example the candidate is mentioning Tom Coburn as a conservative republican who he has work with to justify that he can work with the people of the other party. Thus no propaganda is committed by mentioning the name of this person.
  Example from Mccain-Obama debate, September 26th 2008: Mostly that's just me opposing George Bush's wrong headed policies since I've been in Congress but I think it is that it is also important to recognize I worked with Tom Coburn, the most conservative, one of the most conservative Republicans who John already mentioned to set up what we call a Google for government saying we'll list every dollar of federal spending to make sure that the taxpayer can take a look and see

who, in fact, is promoting some of these spending projects that John's been railing about[9]

5.2.2 Ad hominem

Walton [79] says ad hominem occurs where a statement becomes an excessive personal attack on the candidate's position. Ad hominem propaganda is directed at a person not a situation, administration or strategy, in the case of this data mostly at the other candidate. Ad hominem is an insult to the character or personality not a criticism to what they plan, do or say. [10]

Examples:

- So you've got to ask yourself, why won't he release his tax returns? And I think there may be a couple of reasons. First, maybe he's not as rich as he says he is. Second, maybe he's not as charitable as he claims to be. Third, we don't know all of his business dealings, but we have been told through investigative reporting that he owes about $650 million to Wall Street and foreign banks. Or maybe he doesn't want the American people, all of you watching tonight, to know that he's paid nothing in federal taxes, because the only years that anybody's ever seen were a couple of years when he had to turn them over to state authorities when he was trying to get a casino license, and they showed he didn't pay any federal income tax.[11]

  Justification: With the first statement : "maybe he's not as rich as he say he is" the candidate is implying that the opposing candidate is lying. With the second one they are mentioning that the other candidate is not charitable. These are examples of attacking the person directly thus examples of ad hominem propaganda.

- I'm afraid Senator Obama doesn't understand the difference between

[9] Mccain-Obama debate, September 26th 2008

[10] Boundary Note: General ad hominem propaganda typically has the boundary aligned at sentence-level, or as span-level of a sentence which follow the boundary standard rules.

[11] Clinton-Trump debate,September 26th, 2016

a tactic and a strategy[12]

- Well, I was interested in Senator Obama's reaction to the Russian aggression against Georgia. His first statement was, "Both sides ought to show restraint."Again, a little bit of naivete there. He doesn't understand that Russia committed serious aggression against Georgia[13]
  Justification: Saying that the other candidate has naivete is a ad hominem propaganda.

- Typical politician. All talk, no action. Sounds good, doesn't work. Never going to happen. Our country is suffering because people like Secretary Clinton have made such bad decisions in terms of our jobs and in terms of what's going on[14].
  Justification: The candidate is trying to criticize the person by mentioning stereotypical behaviour of politicians (all action, no talk). Thus, his statement is annotated as ad hominem.

Name calling, labeling

Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable for loves, praises.[15]

Example:

- Manchin says Democrats acted like babies at the SOTU (video) Personal Liberty Poll Exercise your right to vote
  Justification: The use of the expression like babies makes this example an ad hominem because it is trying to diminute their opposite by labeling them as babies.[16]

[12]Mccain-Obama debate September 26th 2008

[13]Mccain-Obama debate September 26th 2008

[14]Clinton-Trump debate, September 26th, 2016:

[15]Boundary Note:

i.) In finding name-calling propaganda, annotator focuses on only a noun-phrase where the words are most evident to evoke negative perceptions.

ii.) An example of a noun-phrase: ([such] + Determiner(the, this, that, a) + Adjective(stupid) + Noun (person)), or a single noun that leads to negatively label a person.

[16]SemEval2020 task-11

Appeal to hypocrisy (tu quoque)

As Walton [79] describes *tu quoque* is a function in dialectic when a candidate tries to evade the statement of the opposite party by "putting the ball in the opponent's court". This technique occurs when a candidate evades the statement of the opposite candidate by saying: *"But you did the same thing!"* instead of justifying their own actions or behaviour. Tu quoque literally means *"you also"*.[17]

Examples:

- The book you mentioned that Vice President Gore wrote, he also called for taxing – big energy taxes in order to clean up the environment.And now that the energy prices are high, I guess he's not advocating those big energy taxes right now.[18]

  Justification: In this example the candidate is pointing out that their opponent is hypocritical about their position on energy taxes and not addressing with a statement why he himself is opposing the energy tax. This is thus an example of *Tu quoque* which is a ad hominem propaganda.

- And I never promoted Fannie Mae. In fact, Senator McCain's campaign chairman's firm was a lobbyist on behalf of Fannie Mae, not me. So – but, look, you're not interested in hearing politicians pointing fingers. [19]

  Justification: This is a very explicit example for *Tu quoquoe* which is a ad hominem propaganda.

Circumstantial ad hominem

This type of propaganda occurs when someone makes a statement by saying that the person making the statement is only making it because it's in their

[17]Boundary Note: The boundaries of tu quoque is typically annotated at sentence-level (one sentence or more).

[18]Bush-Gore debate, October 11th 2000

[19]Mccain-Obama debate, 07 Oct 2008

interest or because of their circumstances. This actually has no bearing on whether or not the statement is true or false.[20]

Example:

- I happen to support that in a way that will actually work to our benefit. But when I look at what you have proposed, you have what is called now the Trump loophole, because it would so advantage you and the business you do.[21]

  Justification: Mentioning that the opposing candidate will take a personal advantage of the tax law they are supporting is an example of circumstantial ad hominem.

Exceptions:

- Criticizing someone's actions, plans or sayings is not ad hominem. General Ad hominem propaganda technique is an insult or attack on the person's character.

- Ad hominem propaganda is concerned with a person, or a group of people. If there is a negative comment or insult toward a system, administration or a non-animate entity the propaganda is more likely to be Loaded Language under the category of Appeal to emotion.

- A positive description of someone is never considered ad hominem (unless said extremely sarcastically which annotator does not look for in this data set).

Examples of exception:

- In the following example the candidate is calling the system rotten not a person so it is not name-calling but loaded language.

  "I don't think there are any villains, but, boy, is the system rotten."

[20] Boundary Note: The boundaries of this propaganda is typically annotated at sentence-level (one sentence or more).

[21] Clinton-Trump debate, September 26th 2016

- The following example is a criticism on the plans of the candidate and not a personal insult thus it should not be considered as ad hominem:

  "I believe the programs that Senator Kennedy advocates will have a tendency to stifle those creative energies, I believe in other words, that his program would lead to the stagnation of the motive power that we need in this country to get progress"

5.2.3 Appeal to emotion

Using emotion to support a statement. If the emotion is sensed by the statement being made , then it is propagandist . There are different categories of emotions that politicians use in their statement.[22]

Appeal to pity, ad Misericordiam

According to Walton [79][23], the definition of appeal to pity is "an appeal to pity may be an evasion of relevant considerations needed to make a decision on the issue. For example, in a criminal trial if the defence attorney bases his whole statement on an appeal to pity, it could be reasonable to criticize his statement as a failure for the guilt or innocence of the opposition.

Example:

- So gun laws are important, no question about it, but so is loving children, and character education classes, and faith-based programs being a part of after-school programs. Some desperate child needs to have somebody put their arm around them and say, we love you[24]

[22]Boundary notes:

i.) Except for the propaganda of loaded language, the rest of subcategories in this category of propaganda are typically at sentence-level (one or more sentence). The loaded language propaganda follows the same pattern as name-calling propaganda.

ii.) Annotator considers the noun-phrase as Loaded language. In the case of the implication does not make the noun-phrase out of the context, One must annotate an entire sentence as loaded language.

[23]page 5

[24]Bush-Gore debate September 26th,2000

Justification: Instead of providing relevant statements against conducting gun laws, the candidate tries to appeal to the emotion of the audience to feel pity for the children who commit shooting in schools. Thus the statement is annotated as appeal to emotion.

Flag waving

Flag waving is a propaganda technique in which the the candidate tries to appeal to a group of people by using statements which contain emotions concerning nation, race, gender, political preference or in general a group, idea or country.

Examples:

- In 1933, Franklin Roosevelt said in his inaugural that this generation of Americans has a rendezvous with destiny.I think our generation of Americans has the same rendezvous. The question now is: Can freedom be maintained under the most severe tack - attack it has ever known? I think it can be. And I think in the final analysis it depends upon what we do here. I think it's time America started moving again [25]

  Justification: By constantly talking of American will, the candidate is trying to appeal to the patriotic emotion to imply that a move (away from the previous administration) is needed in the US. So his statements are identified as appeal to emotion propaganda because it is not explicitly mentioning why this change is needed.

- You know, my father came from Kenya. That's where I get my name. And in the '60s, he wrote letter after letter to come to college here in the United States because the notion was that there was no other country on Earth where you could make it if you tried. The ideals and the values of the United States inspired the entire world. I don't think any of us can say that our standing in the world now, the way children around the world look at the United States, is the same. And part of

[25] Kennedy-Nixon September 26th 1960

what we need to do ,what the next president has to do – and this is part of our judgment, this is part of how we're going to keep America safe – is to – to send a message to the world that we are going to invest in issues like education, we are going to invest in issues that – that relate to how ordinary people are able to live out their dreams[26]

Justification: In this example the candidate is trying to show his competence. The statement he provides using appeal to emotion (flag waving technique) has no relevance to his statement of being competent as the next president.

Appeal to fear

Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative.

Examples:

- Jim, we've got the capability of doing both. As a matter of fact, this is a global effort. We're facing a group of folks who have such hatred in their heart, they'll strike anywhere, with any means.[27] And that's why it's essential that we have strong alliances, and we do. That's why it's essential that we make sure that we keep weapons of mass destruction out of the hands of people like Al Qaida, which we are.

  Justification: The statement that the enemy has hatred in their heart and their attack would be merciless is not a logical statement why alliances are needed in this war, and is used to provoke the audience into being afraid of the enemy anywhere they are. Thus it is a type of appeal to emotion.

- Well, I think it's terrible. If you go with what Hillary is saying, in the ninth month, you can take the baby and rip the baby out of the womb of the mother just prior to the birth of the baby.[28]

  Justification: The candidate is trying to put fear of the law which

[26]Mccain-Obama debate, September 26th 2008
[27]Bush-Kerry debate, September 30th, 2004
[28]Clinton-Trump debate, October 19th 2016

the other candidate is proposing by painting an image of a baby ripped out of their womb as a statement to justify why this abortion law is not good. Thus they are committing an appeal to emotion.

Loaded Language

In this category of propaganda, politicians make use of specific words and phrases with strong emotional implications (either positive or negative) to influence an audience.

In annotating loaded language propaganda , using intensifying adverbs and adjectives which amplify a negative or positive emotion of an expression can be a hint, such as "a tremendous bearing", "a deadly competition".

Examples:

- Well, I actually agree with that. I agree with everything she said. I began this campaign because I was so tired of seeing such foolish things happen to our country[29]

  Justification: The word "foolish" has a negative connotation and is a loaded word which will put the statements used in a loaded language and thus the statement is considered propagandist . It is not ad hominem since it is not directed to an opposite candidate.

- But we have to stop our jobs from being stolen from us We have to stop our companies from leaving the United States and, with it, firing all of their people. All you have to do is take a look at Carrier air conditioning in Indianapolis. They left fired 1,400 people. They're going to Mexico. So many hundreds and hundreds of companies are doing this. We cannot let it happen. Under my plan, I'll be reducing taxes tremendously, from 35 percent to 15 percent for companies, small and big businesses. That's going to be a job creator like we haven't seen since Ronald Reagan. It's going to be a beautiful thing to watch[30]

[29]Clinton-Trump debate, 09 October 2016

[30]Clinton-Trump debate, September 26th 2016

Justification: By using the word "beautiful" the candidate is making a statement that his tax plan is going to be successful using a positive language. Thus committing a propaganda of loaded language.

- to think that another round of resolutions would have caused Saddam Hussein to disarm, disclose, is ludicrous, in my judgment. It just shows a significant difference of opinion [31]
  Justification: Having provided no supporting statements why he thinks the resolutions would not work by using the word ludicrous the candidate is committing a propaganda of loaded language.

- And we have former members of Congress now residing in federal prison because of the evils of this earmarking and pork-barrel spending[32]
  Justification: By using the expression "The evils of" the candidate is committing a propaganda of loaded Language.

5.2.4 False Cause, Post hoc Ergo Propter Hoc

Based on an initiated order of events, Post Hoc is the propaganda technique that provides the conclusion that some event happens as a result of an earlier event. In general, drive to the conclusion that some event is a result of a situation just because it happened at the same time or after. "This propaganda is usually characterized as the statement from correlation to causation" addressed by Walton [79][33]. This propaganda can happen for example when politicians blame the previous administration or a the party of their opponent for something global or general, or demanding credit for a situation which happened during their/their party's time responsible for an office.[34]

Examples:

[31] Bush-Kerry debate, 30th September 2004

[32] Mccain-Obama debate, September 26th 2008

[33] Page 206

[34] Boundary Notes: This boundary is at sentence-level and usually covers more than a sentence. Because this type of propaganda always involves two events as one event being the cause by the other event.

- During the years between World War I and World War II, a great lesson was learned by our military leaders and the people of the United States. The lesson was that in the aftermath of World War I, we kind of turned our backs and left them to their own devices and they brewed up a lot of trouble that quickly became World War II. And acting upon that lesson in the aftermath of our great victory in World War II, we laid down the Marshall Plan, President Truman did.[35]
  Justification: The candidate is implying that the cause of World War II happening after world War I was that American Army left the area of War. In this example, the candidate is implying that the World war II began in the aftermath of world war I because the American's left the field. This is an example of false cause.

- we have to stop the violence. We have to bring back law and order. In a place like Chicago, where thousands of people have been killed, thousands over the last number of years, in fact, almost 4000 have been killed since Barack Obama became president, over almost 4000 people in Chicago have been killed. We have to bring back law and order. Now, whether or not in a place like Chicago you do stop and frisk, which worked very well, Mayor Giuliani is here, worked very well in New York. It brought the crime rate way down.[36]
  Justification: By saying since Barack Obama(the previous president from the opposing party) has been president there has been thousands of killings in Chicago, he is implying that the laws of the administration is a cause of the events. Which is an example of false cause.

Exception:

- A statement does not contain the propaganda of False cause if the candidate explicitly explain the reason why the second event is caused by the first event.

[35]Bush-Gore debate, October 11th 2000
[36]Clinton-Trump debate, September 26th, 2016

Example of exception:

- In the following example the candidate is directly criticizing the previous administration's economic policies and giving a statement why he thinks they have not worked. Thus is not a propaganda of False cause because the candidate is explaining the economic policy of the previous administration has not worked.
  Example from Mccain-Obama debate, September 26th 2008: Now, we also have to recognize that this is a final verdict on eight years of failed economic policies promoted by George Bush, supported by Senator McCain, a theory that basically says that we can shred regulations and consumer protections and give more and more to the most, and somehow prosperity will trickle down. It hasn't worked

5.2.5 Slogans

A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals. It can appear to invoke the excitement and discourage the counter part.[37]

Examples:

- And we can enforce law. But there seems to be a lot of preoccupation on – not certainly only in this debate, but just in general on law. But there's a larger law. Love your neighbor like you would like to be loved yourself. And that's where our society must heading we're going to be a peaceful and prosperous society. [38]

- I know we have to, but this is a classic example of walking the walk and talking the talk. We had an energy bill before the United States Senate. It was festooned with Christmas tree ornament [39]

[37]Boundary notes: The boundary of this propaganda usually occurs both in sentence and span-level where it follows the annotation standard rules.

[38]Example from Bush-Gore debate October 11th 2000

[39]Mccain-Obama debate, September 26th, 2008

- if it doesn't work, then we have strengthened our ability to form alliances to impose the tough sanctions that Senator McCain just mentioned. And when we haven't done it, as in North Korea – let me just take one more example – in North Korea, we cut off talks. They're a member of the axis of evil[40]
  Justification: Using the familiar term axis of evil to talk about North Korea in his statement is a propaganda of using slogans.

5.2.6 Slippery Slope

The slippery slope is to suggest that an unlikely outcome may follow an act. It refers to an extreme event as an assumption or a conclusion that can cause based on facts. In this type of propaganda that is opposing some action based on the fact that some other implausible extreme event may follow by taking this action.[41]

Example:

- Now what do the Chinese Communists want? They don't want just Quemoy and Matsu; they don't want just Formosa; they want the world.[42]

Finally, it must be noticed that there are expressions which are used in political context which if seen out of context could seem to contain some sort of propaganda, however they are terms which are common in the political discussions (especially in the US) as a reference to a political expression. A few of these terms are, for instance, *honest broker*, *trickle-down economics*, *lip service*.

[40]Mccain-Obama debate September 26th 2008

[41]Boundary notes: The boundaries of this propaganda is usually at sentence-level and it can be more than one sentence. Note that one of the sentences is referring to a fact, whereas the other sentence tends to describe an exaggeration of a consequence event which has not happened.

[42]Kennedy-Nixon debate, October 13th 1960:

5.3 Annotation

For the data set creation, annotation started after a training phase of the annotators, where the understanding of propaganda techniques, and the constraints over the boundaries definition were discussed among all annotators. Gold labels were finalized after a reconciliation phase, during which the annotators tried to reach an agreement on the examples causing conflicts in the annotation. The Inter-Annotator Agreement (IAA) was recalculated after each phase of reconciliation. This annotation task has been take over by three annotators. Two of them are PhD students (third year) in computational linguistics. The third annotator has a master degree in linguistics. Note that all annotators are non-native English speakers, but very fluent in English. Hence, through the annotation of the data set, the annotators were relying on the guidelines, and an English-based dictionary to solve any issue due to language barriers.

The U.S. presidential election debates used in this annotation task arranged each debate by year, month, date respectively. Each topic is split into individual sections within a debate. There are 437 sections in total from the debates holding in the following years: 1960, 1976, 1980, 1984, 1988, 1992, 1996, 2000, 2004, 2008, 2012, and 2016. The annotation assignments to the three annotators are distributed equally in each year of a debate regarding the overall number of sections. The annotation is conducted using INCEpTION[43] proposed by Klie et al. [48], a web-based semantic annotation tool. Figure 5.1 shows a screenshot of the INCEpTION tool during the annotation of the political debates with propaganda techniques. The left block of the screenshot is the text representing the speeches from politicians that are transcribed separated in blocks to identify each turn. According to the annotation assignment, each annotator annotated the assigned sections by highlighting the text snippet that is considered as propagandist with respect to the annotation guidelines (Sections 5.1 and 5.2). After highlighting the propagandist snippet, on the right side of the screenshot, an annotator first selects the main propaganda technique, then selecting the sub-category.

[43] https://inception-project.github.io/

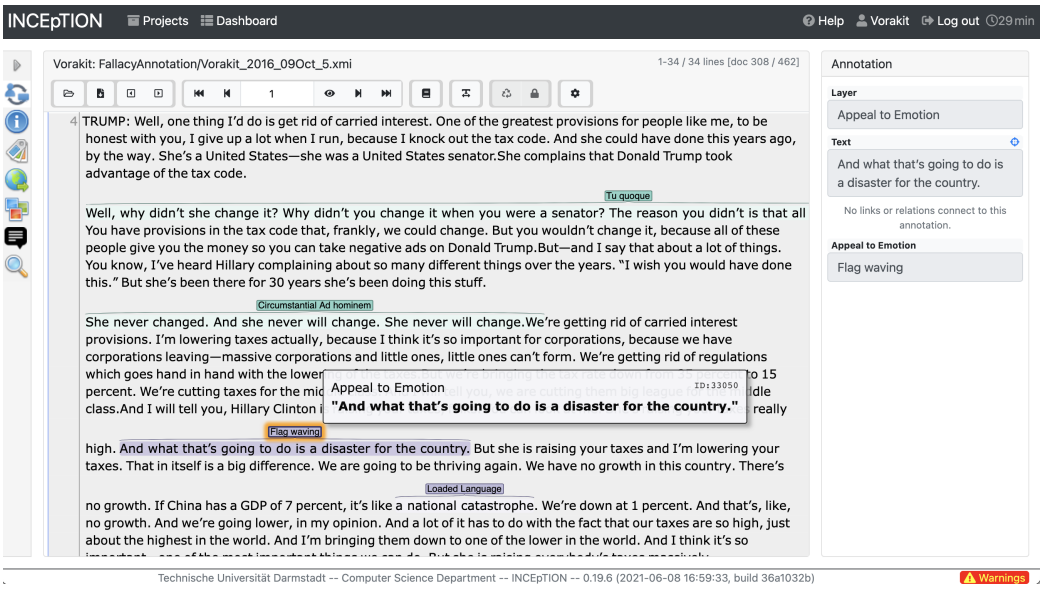


Figure 5.1: A screenshot of the INCEpTION tool during the annotation of the political debate data set with propaganda.

5.4 Evaluation Procedure

Due to the complexity of the annotation task, I compute and report the inter-annotator agreement on different levels, to highlight the difficulty of the task at different granularity levels. Initially, several rounds of discussion among all annotators were held throughout the training process to achieve an agreement on the annotation guidelines. To validate the annotations, the inter-rater reliability or inter-annotator agreement was computed on some, previously unseen, sections of the data after the training period, after which the annotators were familiar with the task and data. First, the pairwise IAA is computed to observe the initial agreement among annotators using sections of multiple debates across the data set after the training process. Then, the last round of discussion ended with the reconciliation phase. Once the obtained IAA was considered as satisfactory, the full annotation phase began. The assigned sections (see Section 5.3 for more details) are then provided to all annotators to separately annotate the data. As the annotation boundary is set differently regarding the propaganda technique to be

annotated, the evaluation score we adopted to compute the IAA was the Krippendorff's coefficient. Given that a severely unbalanced data set may result in instances being correctly classified by chance, using the pairwise IAA only was not sufficient. To consider such sensitive information, we also computed the Krippendorff's α, based on observed disagreement corrected for the expected disagreement, and the Krippendorff's κ (nominal), based on observed agreement corrected for the expected agreement. Both metrics report the same dependability in the case of complete nominal data Zapf et al. [85].

5.4.1 Propaganda Technique Occurrence Agreement

Firstly, I computed whether, over all the sentences in the shared documents among annotators, they agreed on the occurrence of any type of propaganda technique in that sentence. To do so, I exported the annotations from the annotation platform (INCEpTION), then considered the markups (so-called items in the computation of the inter-annotator agreement) to be sentences across all the 10 documents which were mutually annotated by 3 annotators[44]. I used the NLTK tokenize PunktSentenceTokenizer package[45], and setting PunktParameters for considering abbreviations such as: "dr, vs, mr, mrs, prof, inc" which are followed by full-stops to be excluded from sentence rules. Using the toolkit I extracted 1205 sentences out of the 10 shared documents. Thus 1205 items, 2 different categories propagandist-non-propagandist and 3 coders. The terminology is borrowed from Artstein and Poesio [5].

The observed agreement on sentences containing propaganda among the three annotators is 0.9655. The chance corrected inter-annotator agreement measure base don Krippendorff's α is 0.4900. Table 5.1 shows the agreement on sentences containing propaganda between each two pair of annotators

[44]21 Oct 1960 - section 8, 15 Oct 1992 - section 7, 11 Oct 2000 - section 2, 11 Oct 2000 - section 6, 13 Oct 2004 - section 5, 30 Sep 2004 - section 1, 07 Oct 2008 - section 1, 26 Sep 2008 - section 3, 19 Oct 2016 - section 1, 26 Sep 2016 - section 1

[45]https://www.nltk.org/_modules/nltk/tokenize/punkt.html

based on Krippendorff's κ.

A1-A2	A1-A3	A2-A3
0.6382	0.3973	0.4398

Table 5.1: Pairwise Inter-annotator agreement on sentences containing propaganda on the last round of annotation based on Krippendorff's κ nominal.

5.4.2 Inter Annotator Agreement on Propaganda Types

Propaganda Type	Observed Agreement	Chance Corrected Agreement	A1-A2	A1-A3	A2-A3
Ad Hominem	0.9961	0.5315	0.8567	0.3613	0.4978
Appeal to Authority	0.9945	0.5806	0.7680	0.4405	0.5855
Appeal to Emotion	0.9759	0.4640	0.6005	0.3862	0.3940
Slogans	0.9989	0.5995	0.4993	1	0.4993
Overall	0.9914	0.5439	0.6811	0.547	0.4942

Table 5.2: Inter-annotator agreement on different category types on the last round of annotation based on Krippendorff's α for 3 annotators and κ for pairwise agreement.

In this round I compute the inter-annotator agreement on different propaganda types namely: "Appeal to emotion" , "Ad hominem", "Appeal to authority" and the use of "Slogan"s. The two categories of "False cause" and "Slippery Slope" does not have enough samples in the mutually annotated documents to be able to compute the inter-annotator agreement on them. In table 5.2 I report inter annotator agreement for the mentioned categories based on 3 measures. Firstly I report the observed agreement, secondly the chance corrected inter-annotator agreement is reported among the three annotators based on Krippendorff's α measure. Lastly, I report the annotator pairwise chance corrected inter-annotator agreement based on Krippendorff's κ.

5.5 Data Statistics

In this section, I present some statistics on the distribution of the labels associated to the annotated propaganda techniques in the political debate data set.

5.5.1 Statistics by Year of Political Debates

Year of Debate	1960				1976			1980		1984		1988	1992		1996	
Date	07Oct	13Oct	21Oct	26Sep	06Oct	22Oct	23Sep	21Sep	28Oct	07Oct	21Oct	25Sep	13Oct	15Oct	06Oct	09Oct
Annotated Propaganda	34	44	27	49	28	28	11	30	74	28	38	63	45	72	76	69

(a) Annotation in US presidential debates from 1960-1996.

Year of Debate	2000				2004				2008			2012	2016		
Date	03Oct	05Oct	11Oct	17Oct	05Oct	08Oct	13Oct	30Sep	07Oct	15Oct	26Sep	16Oct	09Oct	19Oct	26Sep
Annotated Propaganda	44	53	62	36	47	49	73	63	7	46	49	22	147	153	110

(b) Annotation in US presidential debates from 2000-2016.

Table 5.3: Annotated propaganda in political debates by date and year.

Year of Debate	Ad Hominem	Appeal to Authority	Appeal to Emotion	False Cause	Slippery Slope	Slogans	Total Propaganda
1960	10	24	95	12	12	1	154
1976	5	8	42	4	4	4	67
1980	5	12	77	2	3	5	104
1984	3	15	38	3	3	4	66
1988	4	19	31	2	3	4	63
1996	10	24	93	6	2	10	145
2000	8	25	139	5	8	10	195
2004	32	38	135	13	10	4	232
2008	7	21	67	4	1	2	102
2012	-	2	16	1	1	2	22
2016	98	39	236	9	7	21	410

Table 5.4: Frequency of propaganda categories divided by year.

This section discusses the statistics on the annotated data set of US political debates with propaganda techniques. Table 5.3 refers to the annotation

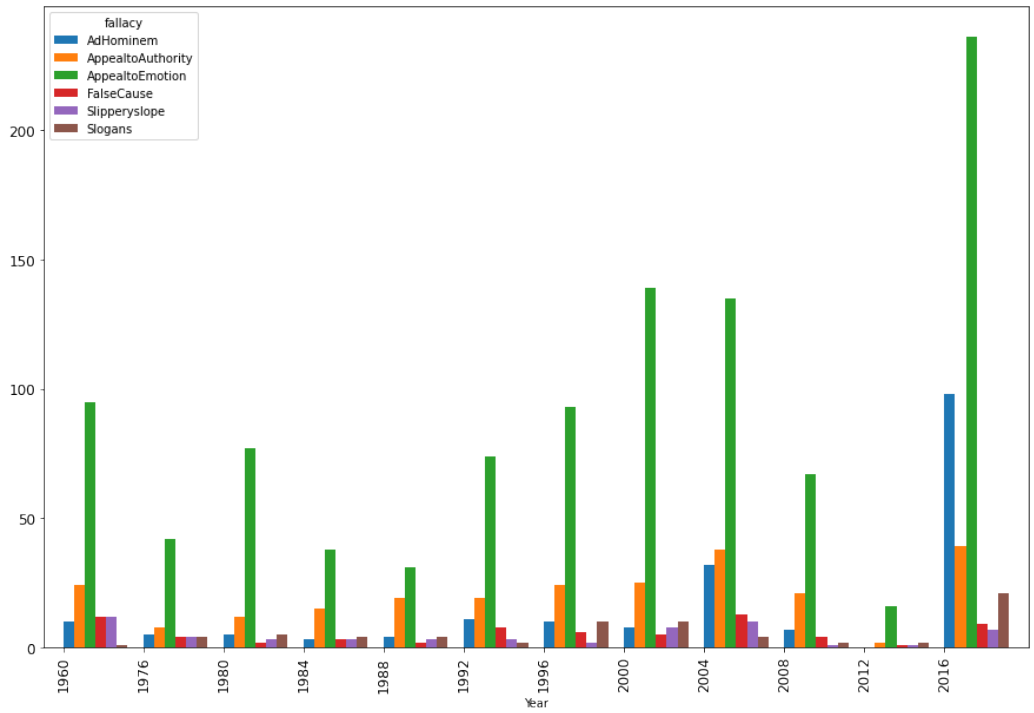


Figure 5.2: A plot showing the frequencies of propaganda annotated in from the data annotation.

based on the year the US presidential debate held. Each date of the debate contains multiple sections (topics). The numbers show the number of propagandist text snippets we detected and annotated for each debate. Throughout the US political debates from year 1960 to 2016, Table 5.4 presents the detailed statistics on the annotated categories of propaganda divided by year. Regarding the year of the debates, Figure 5.2 shows the frequencies of all types of propaganda used by year. The US political debates that contain more propagandist text snippets are in the years 2016, 2004, 2000, 1960, 1996, 1980, 2004, 1976, 1984, 1988, and 2012 respectively. The propaganda technique of "ad hominem" is the most used in the debate of 2016, the "appeal to authority" in 2016 and 2004, the "appeal to emotion" in 2016, the "false cause" in 1960 and 2004, the "slippery slope" in 1960, and "slogans" in 2016. Overall, the most propagandist US political debate we annotated in this dataset is the debate holding in 2016 when Hillary Clinton and Donald

Propaganda Category	Sub-category	Sub-category Frequency	Total
Ad Hominem	Ad hominem	78	190
	Circumstantial Ad hominem	48	
	Name-calling, labeling	35	
	Tu quoque	29	
Appeal to Authority	Appeal to popular opinion	70	245
	False authority	44	
	Without evidence	131	
Appeal to Emotion	Appeal to fear	87	1037
	Appeal to pity	103	
	Flag waving	151	
	Loaded language	696	
False Cause	False cause	69	69
Slippery Slope	Slippery slope	56	56
Slogans	Slogans	69	69
Overall			1666

Table 5.5: Frequency of Propaganda in the US Political Debates from year 1960 to 2016.

Trump were the presidential candidates participating in the debate.

5.5.2 Frequency of Propaganda

Table 5.5 shows the frequency of each propaganda technique annotated on an entire political debate data set. Figure 5.3 highlights the highest frequency which is the "Appeal to emotion" technique (which covers 62% of the data set). Whereas "slippery slope" is the least frequent in the annotated data set (approximately 3% occurrence). Regarding the sub-categories of the propaganda techniques, there are 14 sub-categories. The sub-category "loaded language" in the "appeal to emotion" technique is the most frequent in these US political debates, taking 42% of instances among the annotated data. The sub-category "tu quoque" in the "ad hominem" technique has the lowest occurrence in this data set (only 2% of the whole data set).

As the technique of "appeal to emotion" and "ad hominem" are the most frequent in the annotated data set, Figure 5.4 shows the occurrence of such techniques by year. From 1960 until 1990, in 5.4a and 5.4b both propaganda

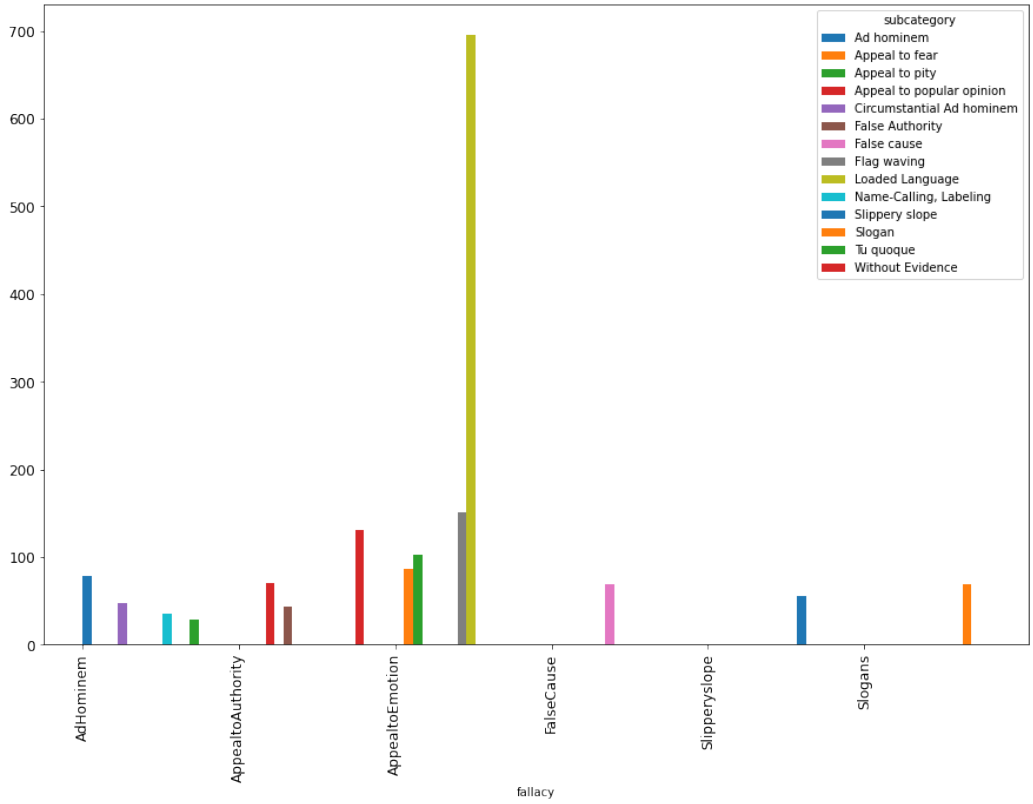
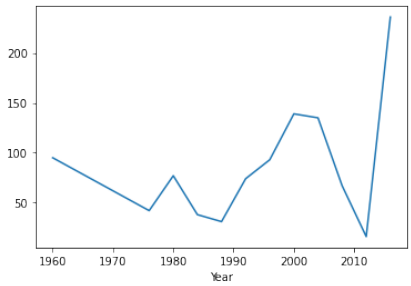


Figure 5.3: A plot showing the frequencies of propaganda annotated in from the data annotation.

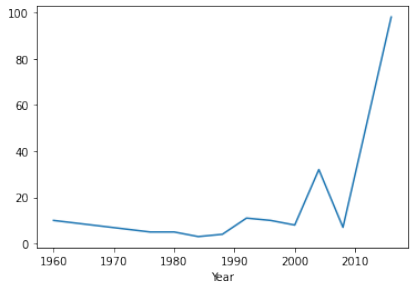
techniques were not used a lot in the US presidential debates. After 1990, the "appeal to emotion" technique has been often employed, more than the "ad hominem" one by 20%. In 2000 until 2008 both propaganda techniques were decreasingly employed in the debates. Later on, in debates from 2008 onward, the "ad hominem" technique has been incrementally used more often. In 2012, both "appeal to emotion" and "ad hominem" were regularly used in 2012, and exponentially used throughout all the debates in year 2016.

5.5.3 Propagandist Text Features in Political Debates

All the propaganda techniques are annotated taking into account the boundary limitations regarding sentence-level and span-level annotations. Figure 5.5 visualizes the word lengths for the annotated data set. Table 5.6 reports



(a) Frequency of appeal to emotion propaganda .



(b) Frequency of ad hominem propaganda .

Figure 5.4: Frequency of most used propaganda over year.

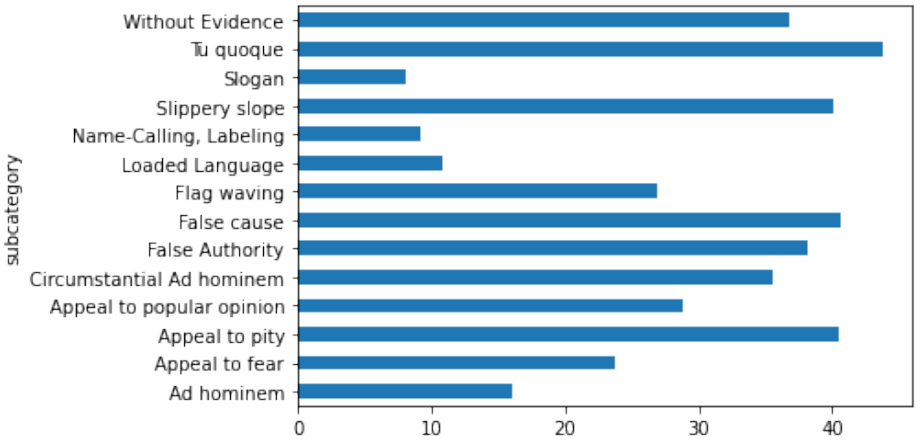


Figure 5.5: A plot showing the average length of propaganda by number of words from the annotated data set.

the details of the average amounts of word counts in descending order for each type of propaganda technique and related sub-categories. The "tu quoque" technique has shown to be the longest in length in these debates. Followed by the "appeal to pity", the "false cause", and the "slippery slope" that have high word counts in the annotated data set respectively. In contrast, the "loaded language" technique, which is one of the most used in these political debates (Figure 5.3), is one of the least word count propaganda technique. The last two in terms of word counts are "name-calling", "labeling" and "slogans" where, as expected, short spans are used to put forward the influence

Sub-category of Propaganda	Avg. Word Count
Tu quoque	44
Appeal to pity	41
False cause	41
Slippery slope	40
False Authority	38
Without Evidence	37
Circumstantial Ad hominem	36
Appeal to popular opinion	29
Flag waving	27
Appeal to fear	24
Ad hominem	16
Loaded Language	11
Name-Calling, Labeling	9
Slogans	8

Table 5.6: Word counts of propaganda by sub-categories

of such propaganda techniques.

5.6 Conclusions

This chapter presents and discusses the collection of a propaganda technique annotated data set on the US presidential debates from 1960 to 2016. All the transcribed debates were segmented by year, month, date and the topic of the debates. Three annotators were involved (including myself) in the training phase to understand the data, the propaganda techniques and the boundary limitations were intensively discussed, followed by a reconciliation phase that allowed to achieve a satisfactory IAA. Six propaganda techniques have been annotated, and they were specified into 14 sub-categories for this annotation task.

The main contribution of this chapter consists in building of a new linguistic resource where propaganda techniques are annotated on the US presidential election debates. This resource provides a new contribution to the

propaganda detection and classification task, with a new reliable annotated resource which will be provided to the scientific community[46] As propaganda detection and classification tasks remain some of the most challenging tasks the NLP community is tackling, the availability of this new data set would be very beneficial for the whole community (see Chapter 2.2.1 for more details on the different data sets released for these tasks), and it could boost the investigation of such propaganda mechanisms in textual data, particularly in the political domain.

[46]The annotated dataset as well as the guidelines will be released upon paper acceptance.

Chapter 6

Proof-of-Concept: the PROTECT System for Propaganda Detection and Classification

This chapter introduces the PROTECT (PROpaganda Text dEteCTion) system, a tool for automatically analyzing text to identify propagandist text snippets and classify them along with the main propaganda techniques. The employment of propaganda in political discourse and debates, and in news articles, as well as its subsequent spread on social networks, may led to threatening consequences for the society and its more vulnerable members. PROTECT is designed as a full pipeline to firstly detect propaganda text snippets from the input text, and then classify the technique of propaganda, taking advantage of semantic and argumentation features. A video demo of the PROTECT system is also proposed to show the main functionalities the user disposes of. This chapter presents the contribution published at the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021) [76].

The goal of propaganda is to persuade the audience about the goodness

of such viewpoint by means of misleading and/or partial arguments, which is particularly harmful for the more vulnerable public in the society (e.g., young or elder people). Therefore the ability to detect the occurrences of propaganda in political discourse and newspaper articles is of main importance, and Natural Language Processing methods and technologies play a main role in this context addressing the propaganda detection and classification task [22; 21]. It is, in particular, important to make this vulnerable public aware of the problem and provide them tools able to raise their awareness and develop their critical thinking.

To achieve this ambitious goal and to design a proof-of-concept for the propaganda detection and classification architecture I described in Chapter 4, I present in this chapter a new tool called PROTECT (PROpaganda Text dEteCTion) to automatically identify and classify propaganda in texts in English. This tool has been designed with an easy-to-access user interface and a web-service API to ensure a wide public use of PROTECT online. To the best of my knowledge, PROTECT is the first online tool for propagandist text identification and classification with an interface allowing the user to submit her own text to be analysed.[1]

PROTECT presents two main functionalities: *i)* the automatic propaganda detection and classification service, which allows the user to paste or upload a text and returns the text where the propagandist text snippets are highlighted in different colors depending on the propaganda technique which is employed, and *ii)* the propaganda word clouds, to show in a easy to catch visualisation the identified propagandist text snippets. PROTECT is deployed as a web-service API, allowing to download the output (the text annotated with the identified propaganda technique) as a `json` file. The PROTECT tool relies on a pipeline architecture to first detect the propaganda text snippets, and second classify the propaganda text snippets with respect to a specific propaganda technique. I cast this task as a sentence-span classification problem and address it relying on a transformer architecture.

[1]The video demonstrating the PROTECT tool is available here `https://1drv.ms/u/s!Ao-qMrhQAfYtkzD69JPAYY3nSFub?e=oUQbxQ`

Results reach SOTA systems performances on the tasks of propaganda detection and classification (the neutral architecture in Chapter 4.4.2 deployed in this pipeline).

This chapter is structured as follows: Section 6.1 presents the dataset used to train and test the pipeline (6.1.1) and a detailed description of the PROTECT architecture (6.1.2). Section 6.2 explains the main functionalities PROTECT serves based on the web. Some concluding remarks end the chapter.

6.1 Propaganda Detection and Classification

PROTECT addresses the task of propaganda technique detection and classification at fragment-level, meaning that both the spans and the type of propaganda technique are identified and highlighted in the input sentences. In the following, I describe the data sets used to train and test PROTECT, and the approach implemented in the system to address the task.

6.1.1 Datasets

To evaluate the approach on which PROTECT relies, I use two standard benchmarks for Propaganda Detection and Classification, namely NLP4IF'19 by Da San Martino et al. [21] and SemEval'20 data sets by Da San Martino et al. [22]. The former was made available for the shared task NLP4IF'19 on fine-grained propaganda detection. 18 propaganda techniques are annotated on 469 articles (293 in the training set, 75 in the development set, and 101 in test set)[2]. As a follow up, in 2020 SemEval proposed a shared task (T11)[3] reducing the number of propaganda categories with respect to NLP4IF'19 (14 categories, 371 articles for train set and 75 in the development set). PROTECT detects and classifies the same list of 14 propaganda techniques as in the SemEval task, namely: *Appeal_to_Authority, Appeal_to_fear-prejudice, Bandwagon, Reductio_ad_hitlerum, Black-and-White_Fallacy, Causal_ Over-*

[2] `https://propaganda.qcri.org/nlp4if-shared-task/`

[3] `https://propaganda.qcri.org/semeval2020-task11/`

simplification, Doubt, Exaggeration_Minimisation, Flag-Waving, Loaded- Language, Name-Calling-Labeling, Thought-terminating_Cliches, Whataboutism-Straw-Men_Red-Herring, Repetition, Slogans,.

Those classes are not uniformly distributed in the data sets. *Loaded-Language* and *Name-Calling_Labeling* are the classes with the higher number of instances (representing respectively 32% and 15% of the propagandistics messages on all above-mentioned datasets). The classes with the lower number of instances are *Whataboutism, Red-Herring, Bandwagon, Straw-Men*, respectively occurring in 1%, 0.87%, 0.29%, 0.23% in NLP4IF'19 datasets. In SemEval'20T11 such labels where merged, and the classes *Whataboutism-Straw-Men-Red-Herring, Bandwagon* respectively represent 1.33% and 1.29% of the propagandist messages.

6.1.2 PROTECT Architecture

Given a textual document or paragraph as input, the system performs two steps. First, it performs a binary classification at token level, to label a token as propagandist or not. Then, it classifies propagandist tokens according to the 14 propaganda categories from SemEval task (T11).

For instance, given the following example *"Manchin says Democrats acted like babies at the SOTU (video) Personal Liberty Poll Exercise your right to vote."* the snippets *"babies"* should be considered as propaganda (step 1), and more specifically an instance of the *Name-Calling_Labeling* propaganda technique (step 2).

Step 1: Propaganda Snippet Detection. I merge the training, development and test sets from NLP4IF, and the training set from Semeval'20 T11 to train PROTECT. The development set from Semeval'20 T11 is instead used to evaluate the system performances.[4] In the preprocessing phase, each sentence is tokenized and tagged with a label per token according to the IOB format.

[4]The gold annotations of Semeval'20 test set are not available, this is why I selected the development set for evaluation.

For the binary classification, I adopt *Pre-trained Language Model* (PLM) based on BERT (*bert-base-uncased* model) [28] architecture. The hyperparameters are a learning rate of 5e-5, a batch of 8, max_len of 128, then a softmax activation function at the prediction layer. For the evaluation, I compute standard classification metrics[5] at the token-level. The resulting macro average over 5 runs on SemEval'20 T11 development set is of 0.72 F1, 0.71 precision and 0.77 recall. I then perform a post-processing step to automatically join tokens labelled with the same propaganda technique into the same textual span.

Given that PLM is applied at token-level, each token is processed into sub-words (e.g., "running" is tokenized and cut into two tokens: "run" and "##ing"). Such sub-words can mislead the classifier. For instance, in the following sentence: "The next day, Biden said, he was informed by Indian press that there were at *least a few Bidens in* India.", my system detects *least a few Bidens in* as a propagandist snippet, but it misclassified one sub-word ("at" was not considered as part of "at least", and therefore excluded from the propagandist snippet).

Step 2: Propaganda Technique Classification. I cast this task as a sentence-span multi-class classification problem. More specifically, both the tokenized sentence and the span are used to feed the transformer[6] as follows: *i)* I input a sentence to the tokenizer where max_length is set to 128 with padding; *ii)* I input the span provided by the propaganda span-template from SemEval T11 dataset, and set max_length value of 20 with padding. If a sentence does not contain propaganda spans, it is labeled as "none-propaganda".

I feed the transformer with the semantic and argumentation features proposed in Chapter 4 into the BiLSTM layer. Such features proved be useful to improve classification performances in propagandist messages classification, obtaining SOTA results on some categories. I apply CrossEntropy as a

[5] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html`

[6] `https://huggingface.co/transformers/`

Propaganda Technique	PLM: RoBERTa
Appeal_to_Authority	0.48
Appeal_to_fear-prejudice	0.57
Bandwagon,Reductio_ad_hit.	0.72
Black-White-Fallacy	0.38
Casual-Oversimplification	0.70
Doubt	0.74
Exaggeration,Minimisation	0.67
Flag-Waving	0.88
Loaded_Language	0.88
Name_Calling,Labeling	0.85
Repetition	0.70
Slogans	0.72
Thought-terminating_Cliches	0.52
Whatab.,Straw_Men,Red_Her.	0.55
Average	0.67

Table 6.1: Results of the PROTECT pipeline for the propagandist text snippet identification and classification tasks on the SemEval'20 T11 development set (macro-F1).

loss function of the BiLSTM as $loss_{\text{proposed\_features}}$, and then add another $loss$ function as follows:

$$loss_{\text{joint\_loss}} = \alpha \times \frac{(loss_{\text{sentence}} + loss_{\text{span}} + loss_{\text{proposed\_features}})}{N_{\text{loss}}}$$

Hyper-parameters are: 256 hidden_size, 1 hidden_layer, drop_out of 0.1 with ReLU function at the last layer before the joint loss function.

To evaluate the PROTECT pipeline to detect propagadist text snippets and then classify them along with the propaganda technique, I merged the data sets of NLP4IF'19 and SemEval'20 T11 (as mentioned in step 1) for the training of the technique classification task. Then I tested PROTECT on the development set from Semeval'20 T11, by compiling the outputs of the snippet detection task (step 1) as a span pattern for the evaluation[7]. Table

[7]This evaluation method follows the evaluation procedure put in place in SemEval'20

6.1 reports on the obtained results averaged over 5 runs (for a full comparison with SOTA systems, see Chapter 4).

Given the high complexity of the task and the classes unbalance, some examples are miss-classified by the system. For instance, in the following sentence "The Mueller probe saw several within Trump's orbit indicted, but not *Trump , as family* or Trump himself", the system annotated the snippet in italics as "Name_Calling,Labeling", while the correct labels would have been "Repetition".

6.2 PROTECT Functionalities

As previously introduced, PROTECT allows a user to input plain text and retrieve the propagandist spans in the message as output by the system. In the current version of the system, two services are provided through the web interface (and the API), described in the following.

6.2.1 Service 1: Propaganda Techniques Classification

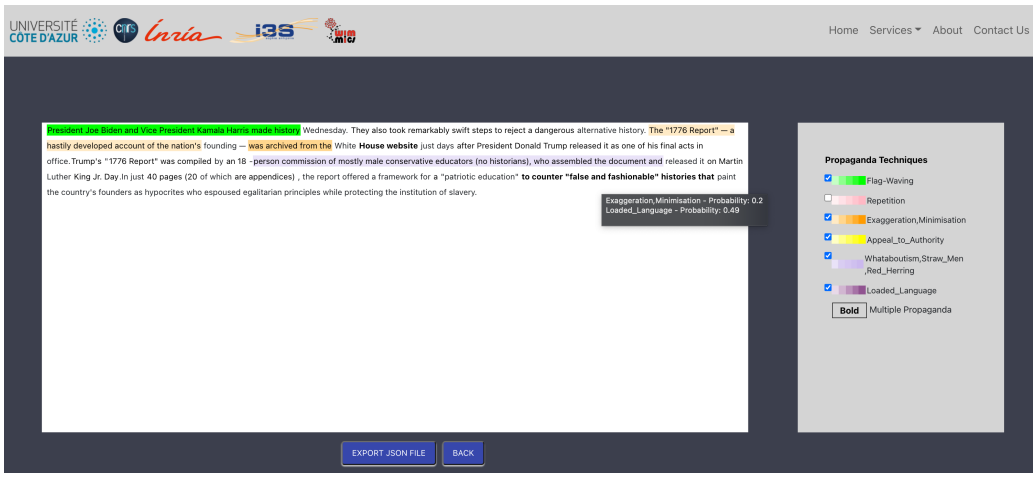


Figure 6.1: PROTECT interface - propaganda techniques classification

T11 for the technique classification evaluation.

The system accepts in input plain text in English, and then the architecture described in Section 6.1.2 is run over such text. The output consist in an annotated version of the input text, where the different propagandist techniques detected by the system are highlighted in different colours. The colour of the highlighted snippet is distinctive of a certain propaganda technique: the darker the color, the higher the confidence score of the system in assigning the label to a textual snippet. Figure 6.1 shows an example of PROTECT web interface. Checkboxes on the right side of the page provide the key to interpret the colors, and allow the user to check or un-check (i.e. highligth or not) the different propagandist snippets in the text, filtering the results. The snippets in bold contain multiple propaganda techniques, that can be unveiled hovering with the mouse over the snippets. The system also allows to download `json` file with the annotated text.

6.2.2 Service 2: Propaganda Word Clouds

Figure 6.2: PROTECT interface - word cloud

The propagandist snippets output by the system can also be displayed as word clouds, where the size of the words represents the system confidence score in assigning the labels (see Figure 6.2). If multiple techniques are found in the same snippet, it is duplicated in the word cloud. As for the first service, a checkbox on the right side of the word clouds allows the user to select the propagandist techniques to be visualized.

6.3 Conclusion

In this chapter, I presented PROTECT, a propaganda detection and classification tool I designed as a proof-of-concept for the classification architecture detailed in Chapter 4. PROTECT relies on a pipeline to detect propaganda snippets from plain text. The rich set of features (i.e., semantic and argumentation features) employed along with SOTA neutral methods deployed make PROTECT a state-of-the-art user-friendly system towards the design of more online disinformation detection tools. I evaluated the proposed pipeline on standard benchmarks achieving state-of-the-art results. PROTECT is deployed as a web-service API that accepts a plain text input, returning downloadable annotated text for further usage. In addition, a propaganda word cloud service allows to gain further insights from such text.

However, disinformation through propaganda and fallacious arguments is not limited to the 14 techniques considered in this thesis, and can be extended to the 24 techniques proposed by Walton [78]. Many factors play then a role, depending on the mean selected by propagandists to spread their views such as government reports, historical revision, books, movies, radio, television, and posters. Propaganda can be found in news, current events, or talk-show segments, as well as in commercial or public service announcement "spots" and long-running advertorials on radio and television. This leaves room for improvement of the PROTECT system to extend the training of the system with information contextualized from heterogeneous propagandist sources, and propagandist text genres.

Chapter 7

Conclusions and Future Perspectives

The information is nowadays transmitted online and in real-time resulting into a growing number of published news from various sources on the Web. In this context, the selection and assessment of the news context and content becomes a challenging and laborious task. This sets the need for systems to (semi-)automatically assist in processing this huge amount of data. The systems for the detection of propagandist texts are to be completely independent and non commercial in order not to be influenced by any commercial considerations. The crucial parts in building such systems are the textual resources annotated for this task, the setting up of optimal classification architectures, and the evaluation settings as the sub-tasks are concerned with fine-grained classification of text snippets.

This thesis addresses three major problems: *(i)* the problem of searching propaganda in online posts. There are several steps contributing to solve this issue, e.g., an observation pipeline to monitor the consistency of the topic content and the audience in order to determine the features to employ in the propaganda detection architectures, a full investigation of the main features that lead to a text to be considered as propagandist; (ii) the demand of structured pipeline systems to detect propaganda and classify it along with the different techniques which may be employed; (iii) the lack of annotated

resources for the propaganda detection and classification tasks, in particular in challenging use case scenarios like political debates.

To address these challenging issues, I proposed the following contributions:

1. **Polarization in Political News Articles and Investigation of Linguistic Distinctive Features of Propaganda** An investigation on public posts is conducted, initially on news political articles in Chapter 3 to observe partisanship or polarization of the audience determined by the source of the newspapers based on the same topic: *"the Brexit"*. The methodology extends the task of aspect-based sentiment analysis and is applied as a tool which tracks the polarization using sentiment elements. The proposed approach showed that, based on sentiment analysis, the elements of polarity (i.e., positive, negative), emotions (e.g., happy, sad, anger, fear), and valence-arousal-dominance hugely helps in capturing polarization in political posts. The proposed approach was able to capture stereotypical aspect-based polarization from newspapers. To better study such techniques from a computational linguistics point of view, an extensive investigation of a set of features that I assumed play a role in propaganda is proposed. Chapter 4 presents a further investigation in representing and extracting the linguistic distinctive features of propagandist texts. In this thesis, the proposed features are investigated to assess how such features help in detecting propagandist texts. These features are persuasion features (e.g., speech style, lexical complexity, concreteness, subjectivity), sentiment features (e.g., polarity, emotions, VAD, connotation, politeness), message simplicity features (e.g., exaggeration, length, pronouns), and argumentation features (e.g., argumentative labels, augmentation components). These features are proved to boost the classification accuracy in all fine-grained tasks ranging from sentence-level to fragment-level (e.g., token- and span-level).

2. **Fine-grained Architectures for the Identification and Classification of Propagandist Texts** The thesis introduced a full

pipeline in propaganda detection and classification with a downloadable outcome analysis of the user text. Chapter 4 reports the results of the experiments with feature-based logistic regression against transformer-based architectures by employing the features discussed in Chapters 3 and 4 to classify sentences into either propaganda or non-propaganda. After getting satisfactory results on the proposed feature representations at sentence-level, I subdivided the propaganda classification tasks into a more fine-grained setting: at fragment-level, where two sub-tasks are described and proposed, firsly, a classification at token-level, and secondly, a classification at span-level. Employing the proposed feature representations in the fine-grained tasks with classical logistic regression, transformer-based architectures, and the proposed ensemble transformer-based architecture have obtained a boost of 10% in average performance with respect to the SOTA on all these fine-grained tasks. This boost in performance affects in particular propaganda techniques like Bandwagon-Reductio-ad-hitlerum, Causal-oversimplification, and Thought-terminating-cliche. Moreover, the ensemble architecture is proposed to tackle particularly the fine-grained task at span-level. Based on such proposed architecture and the satisfactory results I obtained, a full pipeline for propaganda detection and classification is then presented in Chapter 6 for two tasks, namely *(i)* propaganda snippet detection, and *(ii)* propaganda technique classification. The PROTECT pipeline obtains an accuracy in average of 0.67 F1-score for the full pipeline evaluation. For the individual propaganda techniques, the Flag-Waving and the Loaded-Language obtain the highest F1-score of 0.88, which is 10% higher than the best performing SOTA system. The lowest F1-score I obtained is 0.38 F1-score for the Black-White-Fallacy technique.

3. **Creation of a Dataset of Political Debates for Propaganda Detection and Classification** The dataset was built from the transcripts of the United States presidential debates from 1960 until 2016. In the study conducted in Chapter 5 over this annotated linguistic re-

source I built, such political debates tend to contain propaganda techniques such as Flag-waving and Loaded-language. The annotation of propaganda in political debates is addressed in this last part of the thesis. There are 6 types of propaganda techniques peculiar to the domain of political debates, which can be decomposed into 14 further sub-categories. The annotation has been conducted by three annotators who have to determine both the propaganda technique and the sub-category (if any) per an annotated element. The training and consolidation phases of the annotation have been conducted before the actual annotation phase. At the end of the annotation process, 1666 instances of propaganda techniques have been identified in the dataset. In the end, the inter-annotator agreement is computed using $Krippendorff's\alpha$ and pairwise agreements. The reported agreement from all annotators is 0.99 for the observed agreement, and 0.54 for the chance corrected agreement.

In summary, the research presented in this thesis demonstrates how to employ and develop propaganda detection and classification algorithms for the political domain, namely political news articles and political debates, resulting in the annotation of a new dataset for these tasks. Since the field is still emerging, and to support future research in the area of propaganda detection in the political domain, this new dataset on the US political debates will be made available to the community, along with the detection and classification methods described in this thesis. The above listed contributions are valuable input to motivate the community to build upon this work, and reuse this dataset.

Future Perspectives

While important concepts have been carved out in my work, it leaves space for further research directions and future improvements.

First, many different propagandist sources are available online, for instance, news articles, forums, debates, social media channels like Facebook

Pages or Groups, Twitter, etc. As it can be observed from a linguistic point of view, every source has its own characteristics in terms of textual components, styles, and syntax. In the context of this thesis, only political news articles and debates are considered. A first future research direction consists in investigating which features are shared all across these heterogeneous sources.

The investigation of the correlation between sentiment analysis and the proposed features that are encoded in propaganda messages in Chapters 3 and 4, however, together with the virality indexes on social media (e.g., number of likes and retweets, number of replied messages) could help in identifying such propaganda particularly on social media. In Chapter 3, the polarization of texts in news posts is investigated and yielded to the explicit segmentation of audience' groups. The same methods could apply across texts on social media platforms where the text is informal, contains abbreviations, emoticons and slang words, and a higher degree of sentiment and emotions (e.g., leading often to hate speech) which could hint other textual features to help boosting the detection via neural architectures.

Additionally, I aim to improve the performance both of the propaganda detection task and of the classification of propaganda techniques. Several flaws in the utilized Language Model based transformer models have been discovered throughout this thesis, highlighting that current pre-trained models produce outstanding outcomes but are not the solution to all problems. There is room for improvement in the classifiers' performance, notably for the token/sequence classification.

As discussed in Chapter 5, the thesis focuses on six propaganda techniques. Each technique has different boundary limitation, for example, Loaded-language tends to be present in a noun-phase form, whereas False-cause tends to contain two full sentences or more to be considered as a complete propagandist text snippet as this technique needs to cover all the context that refers to multiple events to make a propagandist and fallacious conclusion. Moreover, bias is another concern in this context which directly relate to the annotators toward certain topics regarding race, religion, nation and etc. Since politics is usually a sensitive topic, it could be the case that the anno-

tation is unintentionally biased.

Propaganda detection and classification on political debates remains a challenging task due to the lengthy discourse where facts, opinions, or sometimes sarcastic statements are present. Several baselines rely on standard RNN architectures can find some difficulties in detecting such propagandist text in political debates. Given the nature of political debates, the propagandist text snippet can range from span-level to multiple sentences. BERT [28], the current baseline in transformer-based models, accepts only 512 word tokens at maximum which are not suitable to such distinct characters of the utterance in political debates. The future work goes into the direction of an utterance-level detection and classification task to propose a new architecture to support lengthy utterances.

Furthermore, in Chapter 4, I presented a fragment-level classification where an ensemble architecture is proposed using information produced from transformer models for sentence and span. Given a long utterance, the information about the location where the snippet is found within a fixed length of the model can be compulsive to give the model more information particularly regarding the context of the snippets.

To be able to classify propaganda in such long utterances, there is the necessity to be able to extract information about *i)* the context of the utterance, *ii)* the propaganda snippets and their attention masks, and *iii)* the position embedding of propaganda snippets regarding the length of a utterance. An ensemble architecture must comprise multiple classifiers to capture all the elements listed above. Possible SOTA language models to be investigated are Transformer-XL [25] and Longformer [11], where the length limit of the utterance is less restrictive. However, it requires extensive computational resources and data structure for the task.

Overall, the work described in this thesis addresses only some of the many facets of the propaganda detection and classification issue, particularly in political content. Future work could go into the direction of further integrating other methods to distill even more information at the utterance-level. For instance, more fine-grained contextualized information where the consideration from the utterance-, sentence-, and snippet-level could be beneficial

to provide to the classifier, similarly to what it was done for sentence-span classification in Chapter 4. A framework for propaganda detection and classification for political debates would be beneficial for the qualitative evaluation of disinformation. In particular, future work can go further into the direction of interconnecting the provided utterance context regarding long discourses in favor of political debates for the improvement of the propaganda detection system.

Bibliography

[1] A. F. Agarap. Deep Learning using Rectified Linear Units (ReLU). *arXiv e-prints*, art. arXiv:1803.08375, Mar. 2018.

[2] S. R. Ahmad, M. Z. M. Rodzi, N. S. Shapiei, N. M. M. Yusop, and S. Ismail. A review of feature selection and sentiment analysis technique in issues of propaganda. *International Journal of Advanced Computer Science and Applications*, 10(11), 2019. doi: 10.14569/IJACSA.2019.0101132. URL `http://dx.doi.org/10.14569/IJACSA.2019.0101132`.

[3] J. An, H. Kwak, O. Posegga, and A. Jungherr. Political discussions in homogeneous and cross-cutting communication spaces. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01): 68–79, Jul. 2019. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/3210`.

[4] O. Araque, L. Gatti, J. Staiano, and M. Guerini. Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques. *IEEE Transactions on Affective Computing*, pages 1–1, 2019.

[5] R. Artstein and M. Poesio. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, Dec. 2008. ISSN 0891-2017, 1530-9312. doi: 10.1162/coli.07-034-R2. URL `https://direct.mit.edu/coli/article/34/4/555-596/1999`.

[6] S. Baccianella, A. Esuli, and F. Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In

Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf`.

[7] O. Balalau and R. Horincar. From the stage to the audience: Propaganda on Reddit. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3540–3550, Online, Apr. 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.eacl-main.309`.

[8] R. Balasubramanyan, W. W. Cohen, D. Pierce, and D. P. Redlawsk. Modeling polarizing topics: When do different political communities respond differently to the same news? In J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, editors, *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*. The AAAI Press, 2012. URL `http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4525`.

[9] R. Baly, G. Da San Martino, J. Glass, and P. Nakov. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.404. URL `https://aclanthology.org/2020.emnlp-main.404`.

[10] A. Barrón-Cedeño, I. Jaradat, G. Martino, and P. Nakov. Proppy: Organizing the news based on their propagandistic content. *Information Processing Management*, 56, 05 2019. doi: 10.1016/j.ipm.2019.03.005.

[11] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.

[12] J. A. Berger and K. L. Milkman. What makes online content viral? *Journal of Marketing Research*, pages 192–205, 08 2012.

[13] J. Black. Semantics and ethics of propaganda. *Journal of Mass Media Ethics*, 16(2-3):121–137, 2011. doi: 10.1080/08900523.2001.9679608. URL `https://doi.org/10.1080/08900523.2001.9679608`.

[14] G. Bolsover and P. Howard. Computational propaganda and political big data: Moving toward a more critical research agenda. *Big Data*, 5: 273–276, 12 2017. doi: 10.1089/big.2017.29024.cpr.

[15] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25 1:49–59, 1994.

[16] M. Brysbaert, A. Warriner, and V. Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46, 10 2013. doi: 10.3758/s13428-013-0403-5.

[17] M. Carman, M. Koerber, J. Li, K.-K. R. Choo, and H. Ashman. Manipulating visibility of political and apolitical threads on reddit via score boosting. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 184–190, 2018. doi: 10.1109/TrustCom/BigDataSE.2018.00037.

[18] S. Chakotin. *The rape of the masses : the psychology of totalitarian political propaganda*. Labour Book Service London, 1940.

[19] A. Chernyavskiy, D. Ilvovsky, and P. Nakov. Aschern at SemEval-2020 task 11: It takes three to tango: RoBERTa, CRF, and transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1462–1468, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics. URL `https://aclanthology.org/2020.semeval-1.191`.

[20] R. M. Curnalia. A retrospective on early studies of propaganda and suggestions for reviving the paradigm. *Review of Communication*, 5(4):

237–257, 2005. doi: 10.1080/15358590500420621. URL https://doi.org/10.1080/15358590500420621.

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

[22] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics. URL https://aclanthology.org/2020.semeval-1.186.

[23] G. Da San Martino, S. Cresci, A. Barron-Cedeno, S. Yu, R. Di Pietro, and P. Nakov. A survey on computational propaganda detection. In *Proceedings of 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence (IJCAI-PRICAI2020)*, IJCAI-PRICAI2020, Yokohama, Japan, July 2020.

[24] G. Da San Martino, S. Shaar, Y. Zhang, S. Yu, A. Barrón-Cedeño, and P. 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, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.32. URL https://aclanthology.org/2020.acl-demos.32.

[25] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length

context. *CoRR*, abs/1901.02860, 2019. URL http://arxiv.org/abs/1901.02860.

[26] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria, Aug. 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P13-1025.

[27] D. Demszky, N. Garg, R. Voigt, J. Zou, J. Shapiro, M. Gentzkow, and D. Jurafsky. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2970–3005. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1304. URL https://doi.org/10.18653/v1/n19-1304.

[28] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

[29] J. P. Dillard and M. Pfau. *The Persuasion Handbook: Developments in Theory and Practice*. Sage Publications, Inc., 2009. ISBN 9781412976046.

[30] L. W. Doob. *Public opinion and propaganda / Leonard W. Doob.* Archon Books Hamden, Conn, 2nd ed. edition, 1966.

[31] B. Dowden. Fallacies. 2020.

[32] D. Dubin. The most influential paper gerard salton never wrote. *Libr. Trends*, 52:748–764, 2004.

[33] E. Durmus and C. Cardie. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1094. URL `https://www.aclweb.org/anthology/N18-1094`.

[34] E. Durmus, F. Ladhak, and C. Cardie. The role of pragmatic and discourse context in determining argument impact. *CoRR*, abs/2004.03034, 2020. URL `https://arxiv.org/abs/2004.03034`.

[35] A. S. Edelstein. *Total propaganda: from mass culture to popular culture.* Routledge, 2013.

[36] W. G. Eliasberg. Toward a philosophy of propaganda. *Jewish Social Studies*, 19(1/2):51–63, 1957. ISSN 00216704, 15272028. URL `http://www.jstor.org/stable/4465517`.

[37] J. Fawkes. *Public relations, propaganda and the psychology of persuasion*, pages 195–215. Pearson, 3rd edition, 2014. ISBN 9780273757771. Includes bibliographical references.

[38] S. Feng, J. S. Kang, P. Kuznetsova, and Y. Choi. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1774–1784, Sofia, Bulgaria, Aug. 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P13-1174`.

[39] A. Ferreira Cruz, G. Rocha, and H. Lopes Cardoso. On sentence representations for propaganda detection: From handcrafted features to word embeddings. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 107–112, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5015. URL https://www.aclweb.org/anthology/D19-5015.

[40] R. C. Gilman. The general inquirer: A computer approach to content analysis. philip j. stone , dexter c. dunphy , marshall s. smith , daniel m. ogilvie. *American Journal of Sociology*, 73(5):634–635, 1968. doi: 10.1086/224539. URL https://doi.org/10.1086/224539.

[41] Y. Goldberg. *Neural Network Methods for Natural Language Processing*, volume 10. 2017. doi: 10.2200/S00762ED1V01Y201703HLT037. URL https://doi.org/10.2200/S00762ED1V01Y201703HLT037.

[42] M. Guerini and J. Staiano. Deep feelings: A massive cross-lingual study on the relation between emotions and virality. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 299–305, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450334730. doi: 10.1145/2740908.2743058. URL https://doi.org/10.1145/2740908.2743058.

[43] A. Guimaraes, O. Balalau, E. Terolli, and G. Weikum. Analyzing the traits and anomalies of political discussions on reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01): 205–213, Jul. 2019. URL https://ojs.aaai.org/index.php/ICWSM/article/view/3222.

[44] I. Habernal, H. Wachsmuth, I. Gurevych, and B. Stein. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New

Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1036. URL https://www.aclweb.org/anthology/N18-1036.

[45] S. Haddadan, E. Cabrio, and S. Villata. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In A. Korhonen, D. R. Traum, and L. Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4684–4690. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1463. URL https://doi.org/10.18653/v1/p19-1463.

[46] M. Iyyer, P. Enns, J. Boyd-Graber, and P. Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1105. URL https://aclanthology.org/P14-1105.

[47] D. Jurkiewicz, Ł. Borchmann, I. Kosmala, and F. Graliński. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.semeval-1.187.

[48] J.-C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, and I. Gurevych. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics, June 2018. URL http://tubiblio.ulb.tu-darmstadt.de/106270/.

[49] H. Koppang. Social influence by manipulation: A definition and case of propaganda. *Middle East Critique*, 18:117 – 143, 2009.

[50] S. Lai, K. Liu, S. He, and J. Zhao. How to generate a good word embedding. *IEEE Intelligent Systems*, 31(6):5–14, 2016. doi: 10.1109/MIS.2016.45.

[51] H. D. Lasswell. Propaganda technique in the world war. 1938.

[52] Y. Li, J. Zhang, and B. Yu. An NLP analysis of exaggerated claims in science news. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 106–111, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4219. URL `https://www.aclweb.org/anthology/W17-4219`.

[53] A. Lieto. Cognitive biases for the design of persuasive technologies: Uses, abuses and ethical concerns. 2021.

[54] L. Longpre, E. Durmus, and C. Cardie. Persuasion of the undecided: Language vs. the listener. In *Proceedings of the 6th Workshop on Argument Mining*, pages 167–176, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4519. URL `https://www.aclweb.org/anthology/W19-4519`.

[55] N. Mapes, A. White, R. Medury, and S. Dua. Divisive language and propaganda detection using multi-head attention transformers with deep learning BERT-based language models for binary classification. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 103–106, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5014. URL `https://www.aclweb.org/anthology/D19-5014`.

[56] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.

[57] G. Morio, T. Morishita, H. Ozaki, and T. Miyoshi. Hitachi at SemEval-2020 task 11: An empirical study of pre-trained transformer family for

propaganda detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1739–1748, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics. URL https://aclanthology.org/2020.semeval-1.228.

[58] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

[59] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In M. A. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1202. URL https://doi.org/10.18653/v1/n18-1202.

[60] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[61] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1317. URL https://www.aclweb.org/anthology/D17-1317.

[62] J. Risen and T. Gilovich. *Informal Logical Fallacies.* Cambridge University Press, 2007.

[63] J. Roozenbeek and A. Salvador Palau. I read it on reddit: Exploring the role of online communities in the 2016 us elections news cycle. In G. L. Ciampaglia, A. Mashhadi, and T. Yasseri, editors, *Social Informatics*, pages 192–220, Cham, 2017. Springer International Publishing. ISBN 978-3-319-67256-4.

[64] C. Rovira Kaltwasser. Bringing political psychology into the study of populism. *Philosophical Transactions of the Royal Society B*, 376(1822): 20200148, 2021.

[65] J. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980. doi: 10.1037/h0077714.

[66] G. Salton. Some experiments in the generation of word and document associations. In *Proceedings of the December 4-6, 1962, Fall Joint Computer Conference*, AFIPS '62 (Fall), page 234–250, New York, NY, USA, 1962. Association for Computing Machinery. ISBN 9781450378796. doi: 10.1145/1461518.1461544. URL https://doi.org/10.1145/1461518.1461544.

[67] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, Nov. 1975. ISSN 0001-0782. doi: 10.1145/361219.361220. URL https://doi.org/10.1145/361219.361220.

[68] A. Soliman, J. Hafer, and F. Lemmerich. A characterization of political communities on reddit. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, HT '19, page 259–263, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368858. doi: 10.1145/3342220.3343662. URL https://doi.org/10.1145/3342220.3343662.

[69] C. Stab and I. Gurevych. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland, Aug. 2014. Dublin City University and Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/C14-1142`.

[70] P. Stefanov, K. Darwish, A. Atanasov, and P. Nakov. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 527–537, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.50. URL `https://aclanthology.org/2020.acl-main.50`.

[71] C. Strapparava and R. Mihalcea. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/S07-1013`.

[72] M. Travis. Extracting and networking emotions in extremist propaganda. In *2012 European Intelligence and Security Informatics Conference*, pages 53–59, 2012.

[73] E. Troiano, C. Strapparava, G. Özbal, and S. S. Tekiroğlu. A computational exploration of exaggeration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1367. URL `https://www.aclweb.org/anthology/D18-1367`.

[74] Y. Tsvetkov, L. Boytsov, A. Gershman, E. Nyberg, and C. Dyer. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland,

June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1024. URL https://www.aclweb.org/anthology/P14-1024.

[75] V. Vorakitphan, M. Guerini, E. Cabrio, and S. Villata. Regrexit or not regrexit: Aspect-based sentiment analysis in polarized contexts. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, pages 219–224, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.19. URL https://aclanthology.org/2020.coling-main.19.

[76] V. Vorakitphan, E. Cabrio, and S. Villata. PROTECT: A Pipeline for Propaganda Detection and Classification. In *The Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)*, Milan, Italy, January 2021.

[77] V. Vorakitphan, E. Cabrio, and S. Villata. "Don't discuss": Investigating Semantic and Argumentative Features for Supervised Propagandist Message Detection and Classification. In *Recent Advances in Natural Language Processing (RANLP 2021)*, Varna (Online), Bulgaria, Sept. 2021. URL https://hal.archives-ouvertes.fr/hal-03314797.

[78] D. Walton. *The straw man fallacy*. na, 1996.

[79] D. N. Walton. *Informal Fallacies*. John Benjamins Publishing, Jan. 1987. ISBN 978-90-272-7890-6. Google-Books-ID: LQVCAAAAQBAJ.

[80] A. B. Warriner, V. Kuperman, and M. Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45:1191–1207, 2013.

[81] N. C. Wickramarathna, T. D. Jayasiriwardena, M. Wijesekara, P. B. Munasinghe, and G. U. Ganegoda. A framework to detect twitter platform manipulation and computational propaganda. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 214–219. IEEE, 2020.

[82] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada, Oct. 2005. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/H05-1044`.

[83] S. C. Woolley and P. Howard. Computational propaganda worldwide: Executive summary. 2017.

[84] S. Yoosuf and Y. Yang. Fine-grained propaganda detection with fine-tuned BERT. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 87–91, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5011. URL `https://www.aclweb.org/anthology/D19-5011`.

[85] A. Zapf, S. Castell, L. Morawietz, and A. Karch. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, 16, 2016.

[86] F. Zollmann. Bringing propaganda back into news media studies. *Critical Sociology*, 45(3):329–345, 2019. doi: 10.1177/0896920517731134. URL `https://doi.org/10.1177/0896920517731134`.