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

Technical Report: CCC-22-003

by

Marco Emanuel Casavantes Moreno

Doctoral Advisors:

Dr. Manuel Montes-Y-Gómez,
INAOE, Mexico
Dr. Luis Carlos González Gurrola,
Universidad Autónoma de Chihuahua, Mexico
Dr. Alberto Barrón Cedeño,
Alma Mater Studiorum–Università di Bologna, Italy

Instituto Nacional de Astrofísica, Óptica y Electrónica

April, 2022
Santa María de Tonantzintla, Puebla, CP 72840, Mexico.

Contents

Abstract

Technology has changed the way in which people communicate with each other, giving rise to new services such as social networks. Unfortunately, these services can be used to distribute malicious or manipulative content like fake news and propaganda. This raises several concerns about how easy can be to influence a population and spread disinformation. Current strategies to detect computational propaganda are focused on analyzing their presence in news articles, and haven't reached quite thoroughly other sources of information, such as Twitter.

In this work we are going to focus on the detection of propaganda on Twitter, which has the following challenges: scarcity of labeled data, very short texts, high thematic diversity, among others, although it also presents the opportunity to consider information of context, and relying on previously generated data from news articles. We propose: 1) the construction of a new propaganda corpus from Twitter, 2) a multidimensional analysis with contextual-awareness in the form of bias, geographical origins and metadata; and internal dimensions such as writing style, emotions and topics. As preliminary work, we retrieved more than 20k tweets from propagandist and non-propagandist news sources using distant supervision, we recreated baselines on pre-existing collections of propagandist articles and evaluated the usefulness of news articles to tweets in a cross-domain experimental setting. These results support the idea of developing a classification method specifically tailored for the social media domain.

1 Introduction

This technical report is a document that presents a PhD dissertation proposal titled ”A Multidimensional Analysis of Text for Automated Detection of Computational Propaganda in Twitter”, which was approved on April 7th, 2022.

1.1 Propaganda and its evolution

As stated by [1], in 1939 a group of scholars defined propaganda as “an expression of opinion or action by individuals or groups, deliberately designed to influence opinions or actions of other individuals or groups with reference to predetermined ends”. Computational propaganda is then defined in [2] as “propaganda created or disseminated using computational (technical) means”.
There are two types of computational propaganda: automated and non-automated. The identification of automated accounts (also popularly referred to as bots), used to distribute and deceive users on social networks, is temporarily located at the beginning of the year 2010 [3]. However, it was not until the presidential elections of the United States of America in 2016 that their influence and participation during the campaigns drew attention to its effectiveness. Posts created by bots were retweeted at the same rate as posts from humans. It was even shown that legitimate users could not determine what information had been generated by a human or by a bot. Therefore, in this work we refer to non-automated propaganda as manipulative messages created by genuine sources and accounts, such as particulars, groups or news agencies.

1.1.1 Types of information disorder

According to [4], there are two main kinds of information disorder based on the purpose behind it: on one hand we have misinformation, which includes unintentional falseness such as inaccurate dates, statistics or translations; and on the other hand there's malinformation, deliberately created with an intent to harm, such as changing context, dates or content. A middle ground between these two exist in the form of disinformation (see Figure 1), intentionally false content created with the purpose of causing harm. It is driven by three main interests: making money, gaining political power, or causing disturbance for no reason in particular.

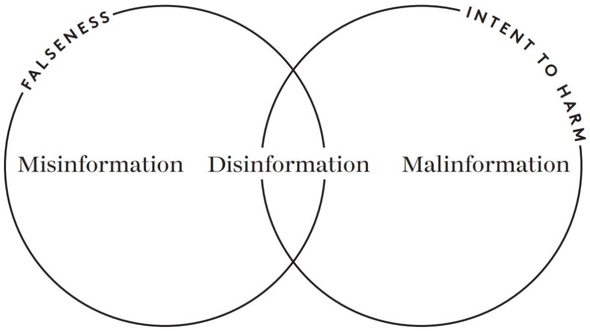


Figure 1: Types of information disorder, borrowed from [4].

We consider important to clarify that propaganda envelops disinformation and malinformation, but since not all propaganda content is generated with bad intentions, some other disorders fit inside the definition of it as a whole, such as Hoaxes (pranking with false stories) or Opinion Spamming (biased reviews towards products or services) [5].

1.2 Challenges and research motivation

The impact that social networks, some of them less than two decades old, has had is considered incredible in terms of the size, scope and speed of growth they achieve every minute, becoming a phenomenon that appears everywhere in our current daily lives [6]. Unfortunately, research indicates that such social networks can also be used to distribute malicious, false or manipulative content [5]. Inside a categorization of this content lies propaganda, which is often associated to news articles and political campaigns promoted in traditional media such as newspapers or websites publishing news as their primary content. However, some research has suggested that, as time changed the resources that people consult and read, social media has also shifted from its traditional use for entertainment to also being an online news provider [7], where the posts are noticeable shorter in length, noisier but easier to digest and anyone can spread a message to thousands of users in a matter of seconds. This raises several concerns about how easy can be to influence a population, and even worse, doing so with the intent to harm. Take for instance the volume of information that was divulged in the 2016 US Presidential campaign aimed to smear the reputation of specific candidates, or the safety and health measures that weren't handed properly at the peak of the COVID-19 global infodemic due to the quantity of disinformation disguised as reliable news [8]. Surprisingly, propaganda detection as a computational task has not been explored as thoroughly as other categories of false information, such as Fake News or Hoaxes [9]. Therefore, there are many aspects of it that remain excluded or isolated in the construction of better detection methods, such as bias levels, geographical origins, metadata, writing style, among others. Each one of these contextual variables represents a different dimension or perspective linked to the propaganda issue. Today's world is in need of automatic tools built to help the struggle that we are living as a consequence of propagandist content created with mal-intent. The aim of this research proposal is to explore propaganda in a social network, compare it to traditional propaganda and tailor strategies to match the shapes and sizes that this content is taking on a social news platform. This study will examine content created in Twitter® by media sources labeled as trustworthy or questionable by their promotion of propaganda.

The remainder of this dissertation proposal is organized as follows. Section 2 introduces a brief discussion about the related work on propaganda detection. Section 3 presents the research proposal that includes the problem statement, research questions, hypothesis, objectives, expected contributions and methodology, along with the work and publication plans. Section 4 contains the preliminary work to support this proposal. Section 5 present the conclusions and Section 6 describes background concepts.

2 Related work

2.1 Propaganda Detection at Document Level

Even tough propaganda has been around since a very long time, it wasn't until 2017 when, in a paper about fake news and political fact-checking [10], propaganda was included in the TSHP-17 dataset to analyze patterns from news articles.

2.1.1 TSHP-17 dataset

To create this corpus, the authors of [10] picked typical trusted news items from the English Gigaword corpus[1] (a large collection of newswire text data in English amassed by the Linguistic Data Consortium over the course of several years), and crawled articles from seven distinct unreliable news sites of various categories (Satire, Hoax and Propaganda). In their study, they investigated linguistic trends across different types of articles, and performed an analytic study characterizing the language of political quotes and news media written with varying intents and degrees of truth. Table 1 shows the quantity of articles in this dataset.

Table 1: News articles in TSHP-17, adapted from [10]

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

[1] https://catalog.ldc.upenn.edu/LDC2003T05

2.1.2 QProp dataset

In 2019, motivated by the difficulties of carrying over further research using the TSHP-17 Corpus due to the small number of propagandist sources and lack of information from individual articles, [11] compiled an improved corpus. This time they considered 94 and 10 sources of non-propaganda and propaganda respectively. Table 2 displays the distribution of their collection. The criteria for labeling news outlets comes from the website MediaBias/FactCheck[2], an online resource that categorizes media according to the bias they exhibit. Their hypothesis was, that classifiers trained with the TSHP-17 Corpus learned to identify the news source because there were only a few of them in the collection. By increasing the size of their corpus selecting more propagandist news sources, future systems trained with this data could learn to distinguish propaganda from texts without such content instead of learning the writing and publishing style of the news outlets. In the paper, a binary class classification was conducted, starting to shape propaganda detection as a standalone task and distancing it further from the fake news scope.

Table 2: News articles in QProp, adapted from [11]

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

2.2 Detection of Propaganda Techniques

2.2.1 PTC Dataset

In 2019, [12] proposed a new dataset with more features that previous collections: to begin with, it was manually annotated instead of using the news source as labels

[2] https://mediabiasfactcheck.com/

(distant supervision), then it was annotated at the span level, meaning that specific snippets of texts were flagged as opposed to full documents. Their last contribution was changing the binary classification scheme for a multi-class classification task, considering 18 propaganda techniques. Although there are some techniques that appear only a few times in the collection (e.g. a technique called “straw man” with 15 instances from a total of 7,485) and therefore may seem unsubstantial, we consider it worth mentioning that the two most popular techniques (appearing 3,841 times combined, more than half the instances in the whole collection) share an association with sentiments and emotions:

- Loaded language.- To affect an audience by using words and phrases with intense emotional connotations (either positive or negative).
- Name calling, labeling.- Using something the target audience either hates or loves to label the object of the propaganda campaign.

As an interesting fact, the authors of [12] now labeled the “trustworthy” class as “non-propagandistic”, perhaps as a result of the difference in task purpose between fake news and propaganda detection. Table 3 shows the distribution of the PTC Corpus.

Table 3: News articles in PTC, adapted from [12]

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

2.2.2 SemEval-2021 Task 6

Pushing for a new modality in detection of persuasion techniques in images and texts, the organizers of [13] used the list of 22 techniques based on previous propaganda research (20 of them applicable to text and 2 to images) to label a collection of english memes from Facebook. The Facebook groups discussed themes such as politics, vaccines and gender equality, resulting in 26 groups crawled over a period of various months in 2020. The annotation step was executed in two phases: 1) independent annotation of memes by annotators, and 2) final gold labels by all annotators and a consolidator. Their final corpus consists of 950 memes, each meme containing at least one persuasion technique.

2.3 Successful approaches at research workshops

Research on computational propaganda has fueled interest in developing solutions to this problem, and some NLP task-oriented workshops have included this area within their activities.

For example, in 2019 the second workshop on NLP for Internet Freedom (NLP4IF) presented two subtasks involving propaganda detection, one for identification of propagandist texts at fragment-level and a binary classification task at sentence-level [14]. In the Sentence-Level Classification, 9 out of 10 teams reported the use of BERT [15] in some form to predict labels, either independently or as part of an ensemble. Other teams from the top scores (shown in Table 4) found useful to consider lexical features, sentiments and tackling the class imbalance of the set to achieve their final results.

Table 4: Top Official Results for NLP4IF SLC Task - Test Set.

Rank	Classifier	F1	System Description
1	BERT	0.6323	Attention Transformer trained on Wikipedia and BookCorpus.
2	BERT	0.6249	Over-sampled training data and performed cost-sensitive classification.
3	BERT	0.6249	Ensemble of models.
4	BERT + LR + CNN	0.6230	Voting ensemble with features from FastText embeddings, readability, emotions and sentiments.
5	N/A	0.6183	Not reported at [14]
6	BERT + USE	0.6138	Ensemble of two BERTs and Universal Sentence Encoder.
7	BERT + bi-LSTM + XGBoost	0.6112	Ensemble with features from GloVe embeddings, affective and lexical representations.

Recent interest in fake news tasks boosted appeal of the detection of propaganda as an active research area. One of SemEval-2020 tasks focused on detection of propaganda techniques in news articles [16], concentrating on fine-grained analysis of texts that could complement existing strategies. Practically all approaches submitted for this task relied on systems based on Transformers. The team with the best score for Span Identification trained several of these architectures and combined them in the end as an ensemble. This result, along with the rest of participants among the top five teams, is displayed in Table 5.

Table 5: Top Results for SemEval-2020 Task 11 Span Identification - Test Set.

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

2.4 Propaganda detection in social media

Authors of [17] explored propaganda from different sources. Their paper hypothesizes that propagandistic sources are sophisticated and creative, and that they will find new ways to deceive by evading trained classifiers. The novelty of their approach lies in cross-domain learning, recognizing the scarcity of labeled data where domains represent different types of sources, such as news articles, social media posts, and public speeches. The data collections used for their experiments fall into precisely these three types of sources. Table 6 shows their distribution of corpora.

1. First, as political speeches, the authors make the contribution of creating a collection of speech transcripts from four politicians, arranged in ordered pairs.

Trump and Obama as contemporary speakers. Trump was seen as more propagandist than Obama. They also use Joseph Gobbels (Nazi Propaganda Minister) and Winston Churchill (UK Prime Minister) as important figures around the time of World War II, Gobbels being more propagandistic than Churchill. All four of these politicians have given propaganda speeches, and their supposition is that two of the speakers exhibit less propaganda than the other two.

2. Second, with news as a source, they combined and reorganized the datasets used in “Hack the News”[3], to build an article-level corpus and a sentence-level corpus.

3. And third, with tweets as a source, they combine two collections, Twitter Russian Internet Research Agency (IRA) from 2016 and considered propagandistic, and twitter7, originally a 2009 collection of almost 476 million tweets from which they took a sample of the same size as Twitter IRA, and that were used as the non-propagandist class.

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

Table 6: Distribution of corpora from [17]

Source	Size (Articles)	Size (Sentences)
Speeches	Trump: 100	Trump: 7,985
	Obama: 100	Obama: 8,336
	Goebbels: 44	Goebbels: 4,482
	Churchill: 44	Churchill: 4,131
	TOTAL: 288	TOTAL: 24,934
News	Propagandistic: 3,899	Propagandistic: 3,938
	Normal: 3,899	Normal: 3,938
	TOTAL: 7,798	TOTAL: 7,876
Tweets	N/A	Twitter IRA: 8,963
		Twitter7: 8,963
		TOTAL: 17,926

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

- Attribute-based models: Logistic Regression and Support Vector Machines. The features considered were word count, weighted n-grams with TF-IDF, and LIWC word categories.
- Models based on neural networks, an LSTM baseline and a modification to this baseline, which is a contribution of this work that they call the LSTM or LSTMR Pairwise Classification Model (seeing that subjective and noisy training labels could lead to over-fitting in traditional supervised learning models, decreasing cross-domain generalization, they designed a model that relaxes the constraints of strict labeling on rankings).

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

The authors of [18] conducted a thorough investigation of propaganda on Reddit. They looked at six political forums in the United States and the United Kingdom, created a dataset (Table 7) and discovered some intriguing patterns:

- Minority parties are more likely to spread propaganda.
- Political leaning might be a sign of propaganda.
- Propaganda techniques in the US are used differently than in the UK.
- Instead of learning topical confounds, their classifier learns the propagandist language pattern.
- Submissions and comments with more propaganda material gain higher engagement, as measured by the number of comments, upvotes, and downvotes.

Their results for propaganda identification are shown in table 8.

Table 7: Reddit dataset distribution, adapted from [18]

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

Table 8: Reddit propaganda identification results, adapted from [18]

Classifier	Precision	Recall	F1
Random	24.14	25.65	28.87
BERT	58.52	52.02	55.08
ROBERTA	63.96	41.41	50.28
XLNet	53.27	59.29	56.12
Ensemble	62.72	48.57	54.74
MGN ReLU	60.41	61.58	60.98

2.5 Discussion

We can see that, when it comes to news articles, there are detection tasks aimed at document level and sentence level. The techniques used to detect propaganda on them mostly involve some kind of transformer-based classifier, either stand-alone or in an ensemble in addition to deep learning models. There exists a study of propaganda on Twitter but using pre-existing collections with labeling schemes based on assumptions, and a Reddit study focused on political forums from the USA and the UK. However, by reading the shortcomings of the related work, we identified some research opportunities:

2.5.1 Scarcity of data and difference of format

The first inclusion of propaganda in the TSHP-17 dataset shows an area of improvement in terms of considered number of propagandist sources, but also inconsistencies in number of documents. For example, although their creators claim to have over 74k articles, their publicly distributed files only account for approx. 39k articles. [11] elaborates on this matter, taking into account more propagandist sources but also describing a more realistic number of documents. Yet, the number of resources aimed specifically towards propaganda detection on social media is still considerably low, not to mention the fact that texts from Twitter are by their nature noisy, they are brief, contain platform-specific features, and are riddled with typos and grammatical errors [19].

2.5.2 Distant Supervision

As noted in a study of related matters about political ideologies [20], a carefully annotated corpus by experts may end up being relatively small, so the authors suggest that future work may explore semi-supervised models or active learning techniques for annotating and preparing larger corpora. Every classifier needs quality data to make good predictions. Similarly to machine learning systems, annotation paradigms can be organized in supervised, unsupervised, and alternative approaches. As part of the latter, the distant supervision scheme, initially conceived for relation extraction purposes. [21], relies on an external database to provide the labeled sources of information to subsequently create instances from them for training data. The labels produced by manual-annotation efforts by experts are considered of higher quality in comparison to distant supervision, however, this paradigm doesn't suffer from some of the disadvantages of hand-labeled supervised approaches, such as being expensive, time consuming and limited in quantity.

2.5.3 Contextual Information

There are different perspectives in the form of contextual information that can be 1)further analyzed to unravel social patterns, and 2)explored and transformed to build a more complete solution to detect propaganda:

- Bias levels and geographic origins: Aside from “non-propagandist” or “propagandist” labels, more dimensions can be associated to news sources, such as their bias levels (from “Extreme-Left” to “Extreme-Right” ideologies) and their country as the place where the news feed is established.

- Topics: As demonstrated in [20, 18], a dependency between propaganda and topics can be studied, this time focusing on dynamics that might be different in other forms of communication such as social media posts.

- Emotions: Some of the most used propaganda techniques are associated with emotions, this suggests that they play an important role in manifestation of propaganda [22, 17].

- Social Media Attributes: In 2020, a survey on computational propaganda detection [8] offered a study focused on tackling this problem from two different perspectives: the content of the propaganda messages and their propagation in social networks. They noted that while there’s research within each one of these aspects, they are isolated from each other and therefore not working together. The authors conclude that, in the near future, it will be necessary not only to take into account propagandist texts but also to analyze the network through which propaganda is disseminated. Most studies perform data collection and subsequent analysis of annotated datasets containing portions of text. However, social media platforms allow their interactions to contain more information aside from the text written in messages. This additional info is called Metadata, it is defined as data that provides information about other data [23].

3 Research Proposal

3.1 Problem Statement

Propaganda can be spread from many different sources, social networks being one of them. The volume of text-based exchanges in social media have made human editorial approaches unfeasible, and recent decisions and rulings by regulatory authorities explicitly mention automatic systems as tools to help mitigate the spread of mischievous content [24], proving their high social relevance.

Shared tasks events are being held online to tackle this challenge and a fair amount of research is published to test new algorithms and approaches. The problem is that most of this research is focused on propaganda extracted exclusively from news articles. To the best of our knowledge, the closest connection to propaganda and social media that we could find comes from [17] as research that combines two pre-existing collections of tweets. Nonetheless, because of the lack of resources and limitations of previous work, the authors of [17] acknowledge the room for improvement and necessity of further research on this subject. To better solve the detection of computational propaganda issue, exploration is needed outside the news articles scope. Since every day the influence of social networks grows as they become the main means of disseminating information, including malicious news and data, the goal of this work is to conduct a multidimensional analysis of computational propaganda, starting from cross-domain strategies to find out if resources from the news articles domain can help to develop a classification system that allows detection of propaganda from social media. Then, our second challenge is to see if propaganda detection can be improved by considering multiple types of context information (such as bias levels, country of origin, emotions used, topic extraction) and modeling that information from multiple dimensions or perspectives.

3.2 Research Questions

The information given in the previous section motivated the proposal of the next research questions:

- How can the resources that exist in the domain of news articles be used to detect computational propaganda in social networks?
- How can contextual information of messages be incorporated to improve the effectiveness of propaganda detection in them?
- How can different ways of modeling the content of messages be considered to improve the effectiveness of propaganda detection in them?
- Should computational propaganda detection be carried out by creating multiple classifiers specialized by thematic content instead of a single-classifier solution?

3.3 Hypothesis

Computational propaganda detection is a complex task. Automatic detection systems are needed to perform in different fields where disinformation is being shared. Our working hypothesis is that resources from the news articles domain can be adapted to detect non-automated computational propaganda in social media posts, and that the combination of content-based and context-based features can be helpful to further improve its detection.

3.4 Objectives

3.4.1 General objective

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

3.4.2 Specific objectives

- To construct a propaganda corpus from social media, retrieving trustworthy and propagandist news sources' tweets.
- To determine the relationship between computational propaganda from news articles and from tweets, proposing a cross-domain strategy to make the most of existing information.
- To propose a propaganda detection model that considers multiple contextual variables such as bias levels, country of origin and metadata.
- To propose a propaganda detection model based on multiple representations of the messages' information such as their content, writing style and emotions.
- To assess the performance of multiple classifiers based on topics against single-classifier solutions to detect computational propaganda.

3.5 Expected Contributions

- The creation of a new corpus of propaganda from Twitter.
- A comprehensive cross-domain analysis of the importance of propaganda articles for the detection of propagandist tweets.

- An approach that incorporates multiple contextual variables for detection of computational propaganda.

- An approach that incorporates multiple representations of messages based on their content for detection of computational propaganda.

- A thematic strategy to adjust the classification method according to the topics included in the texts.

3.6 Methodology

Having established the research questions and the objectives of this work, what follows is to order the necessary steps to fulfill each one of them. This section presents in detail the methodology to reach the proposed objectives. The proposed methodology consists of six stages, where stages 2, 3, 4, 5 and 6 have the major contributions of this dissertation proposal.

Stage 1: Comprehensive study of the state-of-the-art and available resources.

- Obtain previous computational propaganda corpora. As criteria to consider a dataset effective for our study, each collection must be related to propaganda in the form of text posts from propagandist and non-propagandist computational sources.

- Implement state-of-the-art strategies to detect propagandist texts using the aforementioned datasets and set baselines. There are many approaches regarding features to consider and classifiers to use, and experiments putting in practice these strategies based on previous studies will determine how to proceed in our research.

Stage 2: Creation of a new dataset.

This stage involves the creation of a new corpus of computational propaganda content with Twitter as the source of data. The proposed steps are the following:

- Identify sources of propaganda and trustworthy content, labeled as such by MediaBias/FactCheck.
- Create a list comprised of the propagandist and non-propagandist sources that also manage a Twitter account and get their corresponding handle.
- Download the available tweets of the list of sources from two time periods: past (covering a 2017-2018 time frame, same as QProp dataset) and recent (covering 2020-2021).
- Clean the dataset, filtering noisy tweets that do not meet certain criteria such as minimum length and maximum number of hashtags.

Stage 3: Determine the relationship between computational propaganda from news articles and from tweets.

The purpose of this stage is to use the resources that have already been created for the propaganda detection task in news articles. A performance test, training with these data collections, is proposed to determine if they can be adapted to a detection task for the Twitter posts domain. The suggested activities are:

- Analyze the performance of a propaganda classifier trained with articles and tested on Twitter posts to evaluate the relevance of propaganda articles for the detection of propagandist tweets.
- Propose a cross-domain strategy to adapt the classifier from the news to the tweets domain.

Stage 4: Develop a propaganda detection model incorporating multiple contextual information variables.

This stage involves the development of a model architecture to detect computational propaganda adding contextual variables. The following steps are proposed:

- Extract the level of bias from each data source, cluster the sources by geographic origin, and extract platform specific features such as number of favorites, retweets, hashtags and mentions.
- Use the extracted contextual information as new features.
- Use the contextual variables as new dimensions to perform classification on parallel tasks (multi-task learning)

Stage 5: Develop a propaganda detection model extracting information in different representations from the content of the messages.

The following steps are proposed:

- Topic modelling on the short texts to determine their thematic context, analyze the writing style of the news accounts, and perform emotional analysis based on keywords to extract the main emotion associated to each tweet.
- Analyze their results, differences and complementarity.
- Combine the previous representations in a single model.

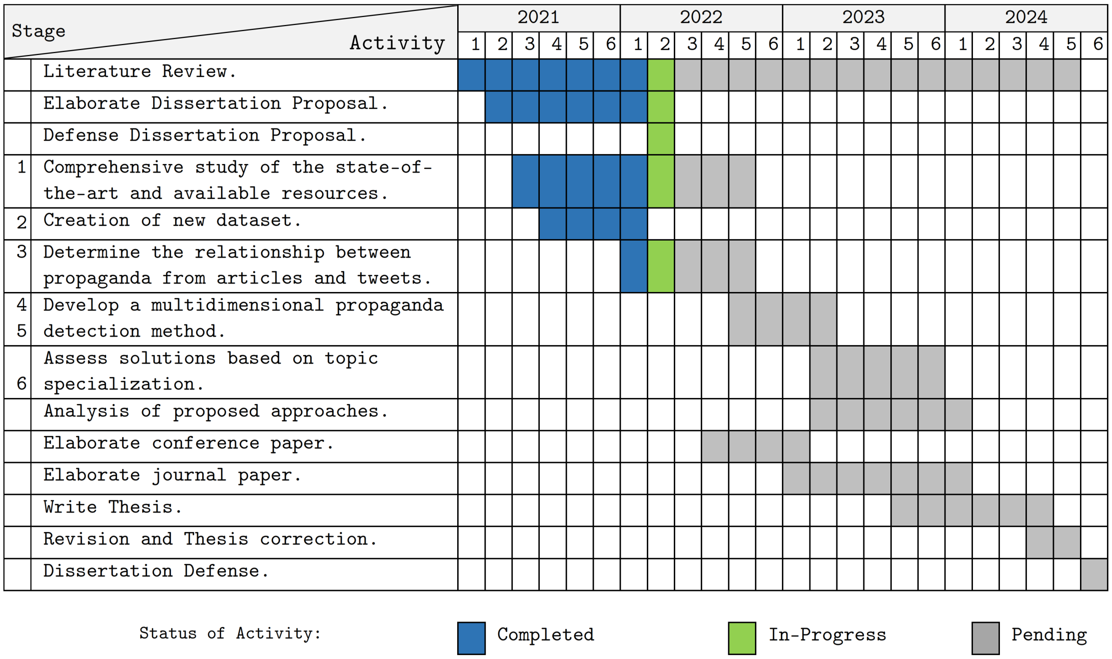
Stage 6: Assess solutions based on topic specialization.

The high diversity of topics covered by news outlets in a data collection present an opportunity. An ensemble of classifiers, each one specialized in a single topic, might be better suited for propaganda detection than a single-classifier solution. In this stage, the following steps are considered:

- Propose a "federated" model for propaganda detection, that considers the combination of several classifiers specialized in particular topics.
- Evaluate and compare the performance of the federated approach against the traditional approach that considers a single classifier.
- Propose a method to adjust the classifier to thematic changes.

3.7 Work Plan

The overall schedule for the time period of 2021-2024 is presented in Figure 2, covering the most relevant activities planned for this research.



Stage	Activity	2021						2022						2023						2024					
		1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
	Literature Review.	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟩	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	
	Elaborate Dissertation Proposal.		🟦	🟦	🟦	🟦	🟦	🟦	🟩																
	Defense Dissertation Proposal.								🟩																
1	Comprehensive study of the state-of-the-art and available resources.			🟦	🟦	🟦	🟦	🟦	🟩	▨	▨	▨													
2	Creation of new dataset.				🟦	🟦	🟦	🟦																	
3	Determine the relationship between propaganda from articles and tweets.							🟦	🟩	▨	▨	▨													
4 5	Develop a multidimensional propaganda detection method.											▨	▨	▨	▨										
6	Assess solutions based on topic specialization.														▨	▨	▨	▨	▨						
	Analysis of proposed approaches.														▨	▨	▨	▨	▨	▨					
	Elaborate conference paper.										▨	▨	▨	▨											
	Elaborate journal paper.													▨	▨	▨	▨	▨	▨	▨					
	Write Thesis.																	▨	▨	▨	▨	▨	▨		
	Revision and Thesis correction.																						▨	▨	
	Dissertation Defense.																								▨

Status of Activity: 🟦 Completed 🟩 In-Progress ▨ Pending

Figure 2: Work plan of activities divided in two-month periods.

3.8 Publications Plan

Table 9 shows a tentative plan of publications.

Table 9: PhD publishing plan.

Type of article	**Target**	**Publication date**	**Content**
For conference	EMNLP	Late 2022	Analysis of differences between propaganda from articles and from tweets.
For journal	LRE	Mid 2023	Description of detection method incorporating context information on new corpus.
For journal	IP&M	Early 2024	Description of detection method incorporating multiple representations of content on new corpus.

4 Preliminary Work

This section presents the preliminary work that has been done to support the feasibility of this research proposal, summarized in the following steps:

1. Identifying and obtaining datasets related to computational propaganda and test current classification strategies (part of the first stage in the methodology).
2. Creating a new dataset of tweets, labeled as propagandist or non-propagandist (part of stage 2 in the methodology).
3. Classifying tweets under a cross-domain setting. A first exploration in using news articles as the auxiliary domain to train classifiers and test them to detect propaganda on tweets as target domain (part of stage 3 in the methodology).

4.1 Obtaining datasets related to computational propaganda.

Our first step on the propaganda detection task is to evaluate current classification strategies. For this purpose, the datasets of news articles from [10, 11, 12] were obtained. These collections are available either by accessing their public repositories or by registering as participants of previous shared tasks for research purposes. Table 10 shows a summary of their sizes.

Table 10: Datasets obtained for Computational Propaganda detection.

Corpus	Level	Documents
TSHP-17 [10]	document	22,580
QProp [11]	document	51,294
PTC [12]	text span	451

The QProp Corpus was selected as the starting point of our experiments since, among the datasets mentioned previously, it shares the most similarities with the kind of data that we want to work with, such as being built using distant supervision, dividing its content into two classes and considering a fair amount of propagandist sources.

4.1.1 Classification of propaganda: experimental Settings

We ran baseline systems with QProp to experiment with different classification models from the main "branches" used on model generalization [25]: traditional baseline, deep learning and transformer-based.

Logistic Regression: classifier with a *lbfgs* solver and $C = 1$. In this experiment, we work with the length of each sentence as the only feature to train, recreating the same baseline from [11].

Convolutional Neural Network (CNN): architecture that mimics the brain's visual cortex in the field of image recognition, but have also proved to be successful in other tasks, such as natural language processing [26]. As hyperparameters we used kernel sizes of [1, 2, 3] and word embeddings of 300 dimensions using the slimmer version of the *word2vec* pre-trained Google News model[4].

Recurrent Neural Network (RNN): a class of networks specialized to work on sequences as inputs, producing an output and then sending it back to itself as a form of memory from previous time steps [26]. The word embeddings utilized are the same as described before for CNN.

BERT classifier: deep bidirectional transformer. We utilized the BERT Base model as feature extractor and classifier, since we are aware it is being used in most NLP tasks achieving state-of-the-art results.

Classifiers and embedding models were implemented from the *scikit-learn* [38], [27], *TensorFlow*, *Simple Transformers* [28] and *Gensim* python libraries [29].

[4] https://github.com/eyaler/word2vec-slim

Table 11: Baselines for Dev set in QProp.

Classifier	F1-score
QProp's MaxEnt [11]	82.93
Logistic Regression	86.88
CNN	90.00
RNN	68.44
BERT	83.80

Table 12: Baselines for Test set in QProp.

Classifier	F1-score
QProp's MaxEnt [11]	82.13
Logistic Regression	85.99
CNN	89.70
RNN	68.81
BERT	84.61

4.1.2 Results

Tables 11 and 12 show the results obtained in the Development and Test sets of QProp, respectively, by the baseline classifiers. After seeing that top results in related workshops were obtained by BERT models, it was unexpected to see a CNN and traditional LR work better for us, but also to see RNN getting the worst performance among them. This indicates that the decision of labeling content as "propaganda" or as "non-propaganda" is being made prioritizing subsets of the input stream instead of focusing on the whole string at once. In other words, the individual terms that are present in the documents are more informative that their sequence itself.

4.2 Creation of a new dataset

The construction of a new data collection began by choosing all available propagandist sources from Media Bias/FactCheck list of questionable sources[5], finding their respective Twitter handle and downloading their tweets as tweet objects from a series of time periods. As subjects, we picked every source of information with the word "Propaganda" as reasoning in their detailed report. We successfully retrieved the tweets shown in Tables 13 and 14, including their date of creation and text content among other metadata features. The tweets were retrieved from two time periods:

- From the beginning of 2021 to the middle of the same year, with almost half

[5] https://mediabiasfactcheck.com/fake-news/

Table 13: Tweets retrieved from 2021-01-01 to 2021-08-20.

Class	Cleaned volume	# of Sources	Avg. tweet length
Non-propaganda	312,143	122	23 words
Propaganda	168,998	124	23 words
Total	481,141	246	

Table 14: Tweets retrieved from 2017-10-01 to 2018-12-31.

Class	Cleaned volume	# of Sources	Avg. tweet length
Non-propaganda	4,429	7	20 words
Propaganda	16,550	33	18 words
Total	20,979	40	

a million tweets from 246 sources, to carry out experiments of stages 4, 5 and 6 from Section 3.6.

- Between 2017 and 2018, to match the collection dates of the QProp dataset [11], so that time periods were not an issue for later cross-domain experiments. Even though we found more than 200 sources of propaganda and non-propaganda combined, the temporal restriction reduced this number to 40.

To clean the new collection of tweets, we discarded retweets and followed guidelines from [30] by removing the tweets that contain three or more trending topics from that time period. We hope that this corpus will be an important contribution to the area of computational propaganda detection, since it would be the first corpus of tweets in English specifically built for this task, and aligned with the current existing ones.

Table 15 presents the top 10 topics extracted from the collection of tweets using Latent Dirichlet Allocation (LDA)[31], where each topic is represented as a multinomial distribution of words (the name of each topic in the first column was designated based on the set of words associated to it).

Table 15: Top 10 topics extracted from tweets dataset.

Topic	Top 10 words (lemmatized and stemmed)
Freedom	live - free - govern - world - nation - today - speech - global- freedom - billion
Middle East 1	saudi - turkey - citi - fact - arabia - death - plan - germani - nation - syrian
Women Rights	right - abort - human - vote - women - democrat - tell - state - life - senat
Religious conflicts	remnant - muslim - polit - islam - confer - christian - america - palestinian - talk - leader
Elections	elect - news - presid - latest - miss - china - fake - lose - mosher - check
US Presidency	trump - want - white - media - obama - peopl - know - liber - presid - russia
General News	publish - newsrescu - school - polic - video - corrupt - children - post - health - student
Middle East 2	buhari - adzw - syria - newspap - nigeria - isi - kill - israel - nigerian - attack
Donald Trump	trump - realdonaldtrump - say - presid - year - countri - peopl - iran - deal - like
Religion	michael - matt - cathol - pope - church - franci - vatican - life - middl - east

4.3 Cross Domain Text Classification

The objective of this experiment consists of testing a propaganda detector under the next scenario: there are available resources from an auxiliary domain in the form of news articles but low resources on a target domain consisting of tweets. We want to determine the label of the tweets, using news articles as the only input for training, and see if the collections of articles are useful to carry out the classification task in the tweets domain.

4.3.1 Preprocessing, Data representation and Experimental settings

All text documents, made up by articles and tweets, were prepared by lowercasing all letters and removing stopwords. After this step, the documents were represented as a Bag-of-words with boolean weighting, to prioritize the importance of presence of terms between domains over their frequency (both domains are contrasting in terms of document length). Our experiments were performed using the following baseline classifiers:

Logistic Regression with a *lbfgs* solver and $C = 1$.

Support Vector Machine with linear kernel and $C = 1$.

Both classifiers were implemented from the *scikit-learn* python library [32].

4.3.2 Results and Analysis

Table 16 shows the in-domain classification results from a 10-fold cross-validation, as well as the scores of the classifier trained with news articles (auxiliary domain) making predictions on the same test partitions of the previous cross validation. We can observe that the classifiers trained out-domain perform far worse than the in-domain equivalent, leading us to believe that the creation of a detection model specifically tailored for tweets from scratch might be a better option.

Table 16: Classification results on tweets dataset.

Model	ACC	PRC	Recall	F1-Score
LR trained with tweets	90.27	87.85	81.52	84.11
Linear SVM trained with tweets	87.95	82.07	81.42	81.73
LR trained with articles	78.82	50.69	50.01	44.21
Linear SVM trained with articles	77.46	43.26	49.34	44.24

To complement the previous experiment, we tried to measure how much the auxiliary domain of news articles is related to the target domain of tweets by using a supervised classifier and a 10 fold cross-validation procedure. Each domain is treated as a class, and each instance is assigned to the class of the domain to which it belongs. If the classifier is able to distinguish between the two distributions, its performance indicates the distance between them [33]. As we can see in Table 17, the results obtained by a Logistic Regression classifier are very high in all the scores, which indicates that the domains are dissimilar, and as a consequence it is easy for the classifier to tell apart a document from another one between domains.

Table 17: Classification results to measure the gap between domains.

Model	ACC	PRC	Recall	F1-Score
LR	99.72	99.69	99.72	99.70

Under these circumstances, there may be relatively few common elements between the article and the tweet domains, and the amount of usable knowledge for training an effective model for a short and noisy domain such as Twitter is minimal.

5 Preliminary conclusions

To conclude this dissertation proposal, we summarize our preliminary conclusions as follows:

- We are creating a new corpus of computational propaganda with Twitter as the source of data. The first part of the collection includes 246 sources and more than 481k of tweets. The second part of the collection includes 40 sources and it has more than 20k of tweets. This represents a contribution to the area of computational propaganda given the scarcity of data from social media to use for research on this particular task.
- Our first experiments in the domain of news articles allowed us to observe that the presence of terms is more relevant than their sequence, and that this allows convolutional neural networks to outperform in some cases other classifiers such as recurrent neural networks and transformers-based at document level.
- Our second experiment, based on a cross-domain text classification approach, allowed us to observe that a cross-domain adaptation seems not so straightforward given the differences between the documents from both domains. A future strategy needs to consider the difference in length and noisiness. We also want to keep testing cross-domain strategies using more classifiers, such as a CNN, a RNN and BERT.

The percentage of completion of this dissertation proposal, based on the Work Plan displayed in Section 3.7 is approximately of **28%**. We are currently focusing our work on Stage 1 and 3 of the Methodology and their respective objectives.

6 Background concepts

6.1 Text Classification

The computational propaganda detection task may be thought of as a text classification problem in which text items are assigned to one or more predetermined groups depending on their content [34]. Every automatic text categorization task involves two major components:

1. Feature extraction through text representations.
2. Machine learning methods of classification.

6.1.1 Text representations

Text data in NLP problems is frequently supplied by human participants and selected from web forums, chat rooms and social media. Machine learning techniques use two prominent feature representations to analyze and extract relevant insights from these texts: *Bag-of-Words* and *Word Embeddings*.

6.1.1.1 Bag-of-Words

Most machine learning applications in the text domain work with the bag-of-words representation (BoW). This model treats each word present in a collection of documents as a feature, and since each file only contains a small subset of the whole vocabulary, BoW is an extremely sparse representation. The value assigned to individual features can be either positive (if the word exists within the document) or zero (if the word is absent). The positive values can be normalized term frequencies or simple binary indicators. For example, consider the next two documents:

- the weenie dog chases a cat

- my cat does not like dry food

A BoW representation of these sentences, filled with binary indicators, would look like Table 18, where each column refers to a term and each row is a document.

Table 18: Example of a *Bag-of-Words*.

	the	**weenie**	**dog**	**chases**	**a**	**cat**	**my**	**does**	**not**	**like**	**dry**	**food**
Doc1	1	1	1	1	1	1	0	0	0	0	0	0
Doc2	0	0	0	0	0	1	1	1	1	1	1	1

Alternatively, a BoW can also consider character n-grams as features (Table 19):

Table 19: Example of a Bag of Character 3-grams.

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

There may be some applications (where a binary input is strictly required, or when presence is more important than frequency) for which binary representations are good enough due to its simplicity. However, if frequency is indeed relevant for the task at hand, the use of normalized frequency of terms is a better way to fill the values of a BoW. This variant is referred to as the *tf-idf* model, where *tf* stands for the *term frequency* and *idf* stands for the *inverse document frequency*. Consider a document collection containing n documents in d dimensions. If $X = (\mathrm{x}_1 \ . \ . \ . \ \mathrm{x}_d)$ is the d-dimensional representation of a document after the term extraction phase, then x_i represents the unnormalized frequency of said document, where all the values of x_i are nonnegative and most are zero [35].

The first step to normalize term frequencies is to compute the inverse document frequency of each term. The inverse document frequency id_i of the ith term is a decreasing function of the number of documents n_i in which it occurs:

$$id_i = \log(n/n_i) \tag{1}$$

The term frequency is normalized by multiplying it with the inverse document frequency:

$$x_i \Leftarrow x_i \cdot id_i \tag{2}$$

One problem with *idf* normalization is that it might increase the frequency of misspellings and errors that weren't handled in the preprocessing stage.

In summary, the universe of words (or terms) corresponds to the dimensions (or features) in this model, turning them into a sparse multidimensional representation, where the ordering of the terms is not used.

6.1.1.2 Word and Document Embeddings

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

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

Clearly, the two sentences are very different but they are identical from the point of view of the bag-of-words representation. For longer segments of text, term frequency usually conveys sufficient evidence to robustly handle simple machine learning decisions like binary classification. This is one of the reasons that sequence information is rarely used in simpler settings like classification. On the other hand, more sophisticated applications with fine-grained nuances require a greater degree of linguistic intelligence. A common approach is to convert text sequences to multidimensional embeddings because of the wide availability of machine learning solutions for multidimensional data. However, the goal is to incorporate the sequential structure of the data within the embedding. Such embeddings can only be created with the use of sequencing information because of its semantic nature [35]. The simplest approach is to use a 2-gram embedding:

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

The linguistic power in the embedding depends almost completely on the type of word-to-word similarity function that is leveraged [35].

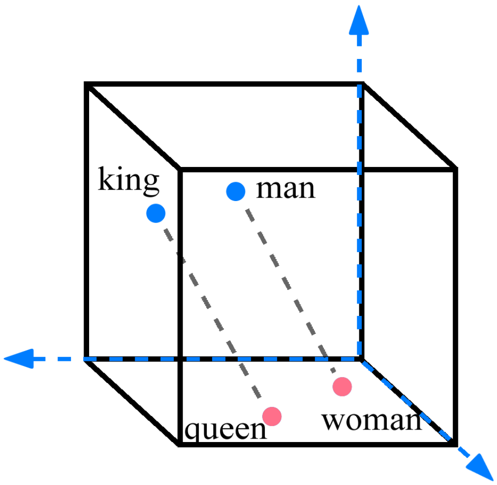


Figure 3: Example of word embeddings on a three-dimensional space.

As shown in Figure 3, the main idea behind this technique is that words that are similar in context (at least according to the text from which the embeddings algorithm trained with) appear closer to each other in a multidimensional space. Based on this, one can use the position of the words in this space to compute the similarity and relation that the text has with its surroundings.

6.1.2 Machine Learning Algorithms

Machine Learning is about making computers modify or adapt their actions (such as making predictions), so that these actions get more accurate, where accuracy is measured by how well the chosen actions reflect the correct ones. It is only over the past decade or so that the inherent multi-disciplinarity of machine learning has been recognized. It merges ideas from neuroscience and biology, statistics, mathematics, and physics, to make computers learn [36]. Machine Learning systems can be classified into broad groups according to the amount and type of supervision they get during training. Some of these categories are: supervised learning, unsupervised learning, semi supervised learning and reinforcement learning [26]. When we feed the training data and the desired solutions or labels to an algorithm, we are talking about supervised learning, and a typical task in this category is classification. The classification problem consists of taking input vectors and deciding which of N classes they belong to, based on training from exemplars of each class. In one-class and multi-class classification problems, each example has one or more labels respectively, but for both tasks the set of classes covers the whole possible output space [36].

For this research, we saw some baselines in related work based on Linear Regression as classifiers. This algorithm is considered part of the traditional approaches for most of the NLP tasks, they are well established, reliable and still competitive to this day.

6.1.2.1 Logistic Regression

Even though it may seem as a contradiction to use the term regression in the name of a classifier, Logistic Regression (LR) is similar to linear regression, with the exception that instead of predicting a continuous value, it simply predicts whether something is true or false, in other words, this algorithm uses a linear regression equation that includes a function called "logistic/sigmoid function", this function produces an "S" shaped curve that is able to tell the probability of class assignment (Figure 4).

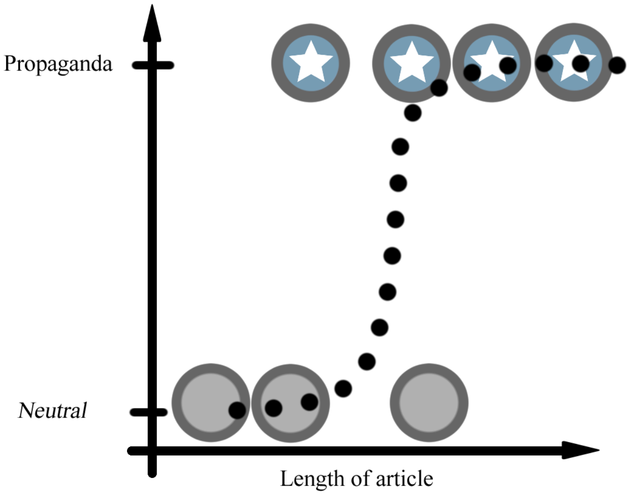


Figure 4: Example of sigmoid function in Logistic Regression.

The sigmoid function is defined [37] as:

$$f(t) = \frac{1}{1 + e^{-t}} \tag{3}$$

Now, we can consider t as a linear function in a univariate regression model [38]:

$$t = \beta_0 + \beta_1 x \tag{4}$$

Therefore, the Logistic Equation becomes:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{5}$$

The choice of the model parameters is a problem that involves finding a hypothesis that best explains our data. The "S" curve is fit to the data using a process called "maximum likelihood". Basically, all the data points are used to calculate the likelihood of the data given the line generated by the sigmoid function. This curve shifts positions until a line with maximum likelihood is selected (see Figure 5).

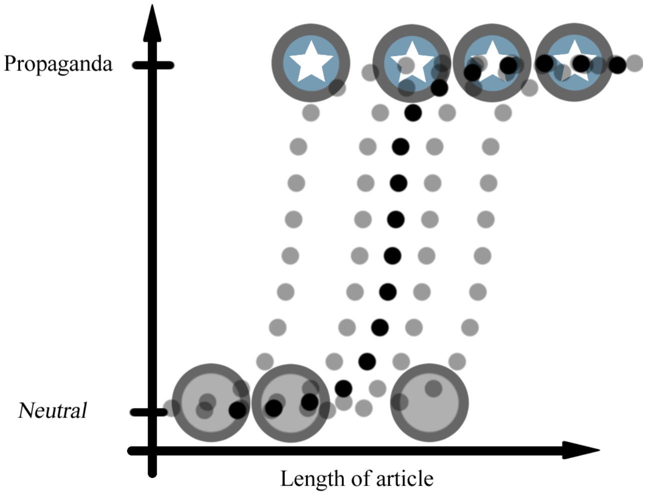


Figure 5: Sigmoid curve tested in different positions to find maximum likelihood.

Overall, this method helps to shrink real valued continuous inputs into a range of *(0,1)* being useful while dealing with probabilities and producing discrete binary outputs [39].

6.1.2.2 Bidirectional Encoder Representations from Transformers

This new representation technique that can also be used to perform classification, better known as BERT by its initials, solves a restriction that current pre-trained language models have, unidirectional architectures. By masking a portion of tokens from the input in a random process called "masked language model", a BERT representation is enabled to combine left and right contexts, generating a deep bidirectional Transformer. BERT's framework consists of a pre-training step, which involves training parameters on unlabeled data, and a fine-tuning step that continues adjusting these parameters, only this time with labeled data from downstream tasks. This process is illustrated in Figure 6 as a question-answering example.

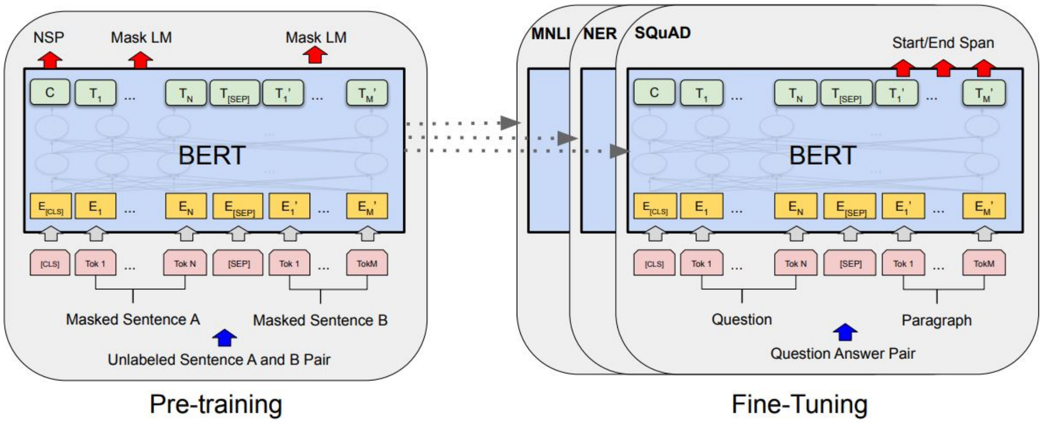
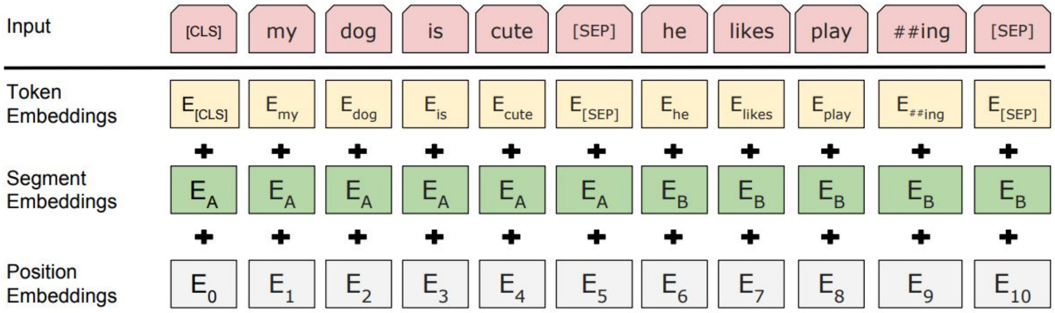


Figure 6: General pre-training and fine-tuning mechanisms in BERT, adopted from [15]. Both pre-training and fine-tuning of parameters use the same architecture.

In BERT, a "sentence" refers to an arbitrary span of adjacent text, and a "sequence" indicates the input token sequence. Each sequence has special tokens, such as "[CLS]" which symbolizes the beginning of the input, and "[SEP]", which separates sentences. The construction of an input representation for a given token, pictured in Figure 7, is the sum of the token, segment and position embeddings.

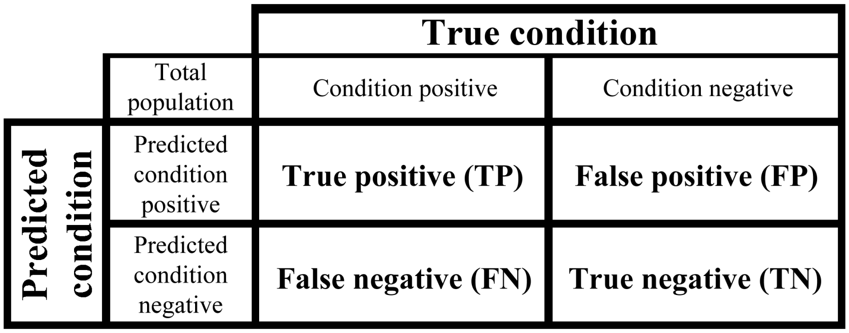


Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

Figure 7: Example of BERT input representation, adopted from [15].

6.2 Evaluation measures

Classification tasks in supervised learning involves comparing predictions against the true labels of samples to train models. The possible outcomes of this comparison is shown in Figure 8.



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

Figure 8: The four outcomes of a 2x2 confusion matrix.

To evaluate the performance of the classifiers, the main classification metrics are considered, as defined in [40]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F1\text{-}score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{9}$$

6.3 Network-based features from Tweet objects

Also known as “status updates”, these objects represents tweets. Each object has a list of fundamental properties. Table 20 displays the tweet attributes and descriptions [41] that are relevant for this research.

Table 20: Metadata of Tweet object.

Attribute	**Type**	**Description**
Retweet count	Integer	“Number of times this Tweet has been retweeted”.
Favorite count	Integer	Indicates approximately how many times the Tweet has been liked by Twitter users.
Replies count	Integer	Number of times the Tweet has been replied to.
Date of creation	Date	“UTC time when this Tweet was created”.

References

[1] C. Miller, *How to Detect and Analyze Propaganda ...: An Address Delivered at Town Hall, Monday, February 20, 1939*. A Town Hall pamphlet, Town Hall, Incorporated, 1939.

[2] G. Bolsover and P. Howard, “Computational propaganda and political big data: Moving toward a more critical research agenda,” *Big data*, vol. 5 4, pp. 273–276, 2017.

[3] G. Bolsover and P. Howard, “Chinese computational propaganda: automation, algorithms and the manipulation of information about chinese politics on twitter and weibo,” *Information, Communication Society*, vol. 22, pp. 1–18, 05 2018.

[4] C. Wardle, “FIRST DRAFT’S Essential Guide to Understanding information disorder,” *First Draft*, no. October, p. 61, 2019.

[5] P. Meel and D. K. Vishwakarma, “Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities,” *Expert Systems with Applications*, vol. 153, p. 112986, 2020.

[6] G. Meikle, *Social Media: Communication, Sharing and Visibility*. Routledge, 2016.

[7] N. Newman, W. Dutton, and G. Blank, “Social media in the changing ecology of news: The fourth and fifth estate in britain,” *International Journal of Internet Science*, vol. 7, 07 2012.

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

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

[10] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, “Truth of varying shades: Analyzing language in fake news and political fact-checking,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (Copenhagen, Denmark), pp. 2931–2937, Association for Computational Linguistics, Sept. 2017.

[11] A. Barrón-Cedeño, I. Jaradat, G. Da San Martino, and P. Nakov, “Proppy: Organizing the news based on their propagandistic content,” *Information Processing & Management*, vol. 56, 05 2019.

[12] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov, “Fine-grained analysis of propaganda in news article,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (Hong Kong, China), pp. 5636–5646, Association for Computational Linguistics, Nov. 2019.

[13] D. Dimitrov, B. B. Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, and G. Da San Martino, “Semeval-2021 task 6: Detection of persuasion techniques in texts and images,” in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pp. 70–98, 2021.

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

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long*

and Short Papers), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

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

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

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

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

[20] B. M. Sinno, B. Oviedo, K. Atwell, M. Alikhani, and J. J. Li, "Political ideology and polarization of policy positions: A multi-dimensional approach," *ArXiv*, vol. abs/2106.14387, 2021.

[21] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, (Suntec, Singapore), pp. 1003–1011, Association for Computational Linguistics, Aug. 2009.

[22] Y. Hua, "Understanding BERT performance in propaganda analysis," pp. 135–138, 2019.

[23] "Metadata — Definition of Metadata by Merriam-Webster." `https://www.merriam-webster.com/dictionary/metadata`. (Accessed on 2020-01-30).

[24] A. Sardo, "Categories, balancing, and fake news: The jurisprudence of the european court of human rights," *Canadian Journal of Law & Jurisprudence*, vol. 33, no. 2, p. 435–460, 2020.

[25] P. Fortuna, J. Soler-Company, and L. Wanner, "How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?," *Information Processing & Management*, vol. 58, no. 3, p. 102524, 2021.

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

[27] R. Rehurek and P. Sojka, "Gensim–python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.

[28] T. C. Rajapakse, "Simple transformers." `https://github.com/ThilinaRajapakse/simpletransformers`, 2019. (Accessed on 2021-04-22).

[29] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[30] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pp. 591–600, 2010.

[31] J. C. Campbell, A. Hindle, and E. Stroulia, "Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data," *The Art and Science of Analyzing Software Data*, vol. 3, pp. 139–159, 2015.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[33] L. Candillier and V. Lemaire, "Design and analysis of the nomao challenge active learning in the real-world," 08 2013.

[34] K. Aas and L. Eikvil, "Text categorisation: A survey.," 1999.

[35] C. Aggarwal, *Machine Learning for Text*. Springer International Publishing, 2018.

[36] S. Marsland, *Machine Learning: An Algorithmic Perspective*. Chapman and Hall/CRC, 2nd ed. ed., 2014.

[37] T. Mitchell, *Machine Learning*. McGraw-Hill International Editions, McGraw-Hill, 1997.

[38] "Univariate Linear Regression - MuPAD." `https://www.mathworks.com/help/symbolic/mupad_ug/univariate-linear-regression.html`. (Accessed on 2020-02-18).

[39] "LOGISTIC REGRESSION CLASSIFIER - Towards Data Science." https://towardsdatascience.com/logistic-regression-classifier-8583e0c3cf9. (Accessed on 2019-12-02).

[40] D. Olson and D. Delen, *Advanced Data Mining Techniques*. Springer Berlin Heidelberg, 2008.

[41] "Tweet object — Twitter Developers." https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object. (Accessed on 2019-12-02).