Participation

Techniques and Services for Monitoring Recruitment and Propaganda

Deliverable D4.3

Óscar Araque[1], Carlos A. Iglesias[1], Pablo Real[1], Kevin McDonald[2], Necla Acik[2], Lorenzo Marinone[3]

1 UNIVERSIDAD POLITÉCNICA DE MADRID (UPM)

2 MIDDLESEX UNIVERSITY LONDON (MDX)

3 EUROPEAN FOUNDATION FOR DEMOCRACY (EFD)

This project has received funding from the European Union's Horizon 2020 research
and innovation programme under grant agreement No 962547.

Deliverable information

Grant Agreement No.	**962547**
Project acronym	PARTICIPATION
Project title	Analyzing and Preventing Extremism via Participation
Project timeframe and duration	1.12.2020–30.11.2023 (36 months)
WP	*WP4 - CONTRASTING RADICALISATION AND EXTREMISM VIA COMMUNICATION*
Task	T4.3
Deliverable	*D4.3 Techniques and Services for Identifying Recruitment and Propaganda*
Status	*First draft*
Version number	0.6
Deliverable responsible	*MDX*
Dissemination level	Public
Due date	M17
Date of submission	30/04/2022

Project coordinator

Name	Prof. Francesco Antonelli
Organization	Università degli Studi "Roma Tre"
Email	Francesco.antonelli@uniroma3.it
Postal address:	Dipartimento di Scienze Politiche Università degli Studi Roma Tre Via G. Chiabrera, 199 00145 – Rome (RM), Italy

Version history

Version	Date	Author	Description
0.0	1-10-2021	Carlos A. Iglesias (UPM)	Adapt template to deliverable specifics and prepare structure of the document.
0.1	15-12-2021	Carlos A. Iglesias (UPM)	Background section about computational techniques and datasets for propaganda detection.
0.2	20-12-2021	Óscar Araque (UPM)	Review of Background section.
0.3	19-04-2022	Necla Acik (MDX), Kevin McDonald (MDX), Lorenzo Marinone (EFD)	Discussion about case studies and provision of keywords.

0.4	19-04-2022	Óscar Araque (UPM)	Added sections for describing the architecture, propaganda classification based on machine learning and use cases.
0.5	23-04-2022	Óscar Araque (UPM), Carlos Á. Iglesias (UPM) and Pablo Real (UPM)	Added demo and analysis sections.
0.6	25-04-2022	Kevin McDonald (MDX), Óscar Araque (UPM) and Carlos Á. Iglesias (UPM)	Final review
0.7	07-05-2024	Óscar Araque (UPM) and Carlos Á. Iglesias (UPM)	Added section 4.4 of manual annotation and evaluation

Author list

Name	**Organisation**
Oscar Araque	*Universidad Politécnica de Madrid (UPM)*
Carlos A. Iglesias	*Universidad Politécnica de Madrid (UPM)*
Pablo Real	*Universidad Politécnica de Madrid (UPM)*
Necla Acik	*Middlesex University London (MDX)*
Kevin McDonald	*Middlesex University London (MDX)*
Lorenzo Marinone	*European Foundation for Democracy (EFD)*

Table of Contents

List of abbreviations

Acronym	Description
IS	Islamic State
MBFC	Media Bias / Fact Check
NLP	Natural Language Processing
SNA	Social Network Analysis
UNODC	United Nations Office on Drugs and Crime

1. Introduction

According to United Nations Office on Drugs and Crime (UNODC) (Barrett, 2012), the use of the Internet to promote and support terrorist acts can be classified into six overlapping categories: propaganda (including recruitment, radicalization, and incitement to terrorism); financing; training; planning; executing; and cyber-attacks.

Propaganda usually uses multimedia communications to provide ideological or practical instruction, explanations, justifications, or promotion of terrorist activities (Barrett, 2012). While propaganda is generally not forbidden, since the right to freedom of expression protects it, terrorist propaganda usually encourages to perform violent acts through extremist rhetoric. Terrorist propaganda can be focused on recruitment, radicalization, and incitement, which can be viewed as a continuum.

This deliverable is focused on extending the work developed in T4.2 with two main goals: i) enabling the analysis of propaganda and recruitment communications and ii) adding new analysis dimensions to the intelligent engine, such as moral values and emotions. The deliverable is structured as follows. The deliverable is structured as follows. First, Sect. 2 provides a review of the literature of the works that previously addressed these topics, intending to understand the techniques used and the available datasets. Then Sect. 3 describes how the software architecture of the monitoring service developed in T4.2 has been extended for classifying propaganda. In particular, two tools have been developed: a dashboard and a Chrome plugin. First, the monitoring dashboard has been extended for monitoring propaganda and analyzing two specific use cases. In addition, a Participation Chrome Plugin has been developed to make more accesible the adoption of the services in the Participation laboratories and increase the impact of the project. Next, Sect. 4 presents the machine-learning methods developed for classifying propaganda texts based on deep learning techniques. Then Sect. 5 describes the use of case studies that analyze the extremist group "Islamogram" on the social network Reddit and the Italian far-right movement so-called "Mattonisti" on Twitter. Finally, Sect. 6 concludes with a discussion.

2. Background

Computational Propaganda Detection (Martino et al., 2020) is a new research field that has emerged in the Natural Language Processing (NLP) research community, motivated by the recent concern about disinformation campaigns during political elections at the COVID-19 pandemic. While disinformation diffuses false claims, propaganda techniques aim at influencing the opinion of others by diffusing false or truth claims. Therefore, propaganda research is closely related to disinformation research. According to Volkova et al. (2019), deceptive news can be classified into the following types:

- *Misinformation: which contains information created intentionally to mislead the audience.*
- *Propaganda: its primary purpose is influencing, manipulating, and accepting opinions and attitudes.*
- *Hoaxes: include scams or deliberately false or misleading stories.*
- *Conspiracies: forms of explaining an event by reference to the machinations of influential people.*
- *Clickbait: content created to attract web traffic.*
- *Satire: statements whose primary purpose is to entertain.*

From a computational point of view, we can distinguish two main perspectives for analyzing propaganda (Martino et al., 2020): network analysis and text analysis. The network analysis perspective applies mainly Social Network Analysis (SNA) techniques for analyzing the social context in which these messages are interchanged (Sánchez-Rada & Iglesias, 2019). This perspective aims at detecting groups of accounts that exhibit inauthentic coordination behavior. The text analysis perspective applies NLP techniques for identifying propaganda based on the linguistic patterns used in the texts, both news and posted messages on social networks. In the context of the PARTICIPATION project, we will focus on the second perspective since the first perspective is focused on detecting anomalous groups of accounts that are coordinated to propagate propaganda. Instead, the PARTICIPATION project is focused on analyzing the communications approaches, which are addressed from the text perspective. Two main approaches have been followed in the text analysis perspective (Martino et al., 2019): sentence-level and fine-grained level. The sentence-level analysis considers a binary classification task where a sentence is classified as propaganda or not. The fine-grained level aims to detect the span of text where a propaganda technique is used and then follows a multi-class classification problem to classify the detected propaganda technique. The SemEval shared task for "Detection of Propaganda Techniques in News Articles" provides the most well-known classification of propaganda techniques (Martino et al., 2019):

1. **Presenting Irrelevant Data (Red Herring):** Introducing irrelevant material to the discussed issue so that everyone's attention is diverted away from the points made.
2. **Misrepresentation of Someone's Position (Straw Man)**: when an opponent's proposition is substituted with a similar one which is then refuted in place of the original proposition.
3. **Whataboutism**: A technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.
4. **Causal Oversimplification**: Assuming a single cause or reason when there are multiple causes for an issue. It includes transferring blame to one person or group without investigating the issue's complexities.
5. **Obfuscation, Intentional vagueness, Confusion**: Using words which are deliberately not transparent so that the audience may have its interpretations. For example, when a vague phrase with multiple definitions is used within the argument, it does not support the conclusion.
6. **Appeal to authority**: Stating that a claim is accurate simply because a proper authority or expert on the issue said it was true, without any other supporting evidence offered. We consider the particular case in which the reference is not an authority or an expert in this technique, although it is referred to as Testimonial in literature.

7. **Blackandwhite Fallacy, Dictatorship**: Presenting two alternative options as the only possibilities when more possibilities exist. As an extreme case, tell the audience exactly what actions to take, eliminating any other possible choices (Dictatorship).

8. **Name-calling or labeling**: Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable, or loves, praises.

9. **Loaded Language**: Using specific words and phrases with solid emotional implications (positive or negative) to influence an audience.

10. **Exaggeration or Minimisation**: Either representing something in an excessive manner: making things more significant, better, worse (e.g., ”the best of the best”, ”quality guaranteed”) or making something seem less important or more minor than it is (e.g., saying that an insult was just a joke).

11. **Flag-waving**: Playing on strong national feeling (or to any group, e.g., race, gender, political preference) to justify or promote an action or idea.

12. **Doubt**: Questioning the credibility of someone or something.

13. **Appeal to fear/prejudice**: Seeking to build support for an idea by instilling anxiety or panic in the population towards an alternative. In some cases, the support is built based on preconceived judgments.

14. **Slogans**: A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals.

15. **Thought-terminating cliché**: Words or phrases that discourage critical thought and meaningful discussion about a given topic. They are typically short, generic sentences that offer seemingly simple answers to complex questions or distract attention away from other lines of thought.

16. **Bandwagon**: Attempting to persuade the target audience to join in and take the course of action because “everyone else is taking the same action”.

17. **Reductio ad hitlerum**: Persuading an audience to disapprove of an action or idea by suggesting that the idea is popular with groups hated in contempt by the target audience. It can refer to any person or concept with a negative connotation.

18. **Repetition**: Repeating the same message repeatedly so that the audience will eventually accept it.

A recent area of research is the identification of multimedia propaganda (Alam et al., 2021). Several studies have outlined the role of visual propaganda in social networks. To cite a few, Seo (2014) analyzed the different characteristics of visual propaganda posted on Twitter during the 2012 Israeli-Hamas conflict. In addition, Dimitrov et al. (2021) have developed a multimedia annotated dataset extracted from Facebook that classifies memes using 22 propaganda techniques.

Desk research has been carried out to identify relevant datasets for detecting propaganda and recruitment, shown in Table 1.

Several surveys provide an introduction to online extremism detection (Gaikwad et al., 2021a), propaganda detection (Martino et al., 2019), and multimodal disinformation detection (Alam et al., 2021), that provide a good overview of the topic and help to find relevant research works. In addition, D'Ulizia et al. (2021) provides an excellent repository of datasets for fake news detection.

Several authors (Saini & Bansal, 2021; Scanlon & Gerber, 2014) have annotated messages from Dark Web forums to detect recruitment messages. Unfortunately, the annotations of these works are not available.

A. Johnston & Marku (2020) propose a deep learning model developed for identifying extremism in forums. The work uses several (non-available) datasets:

- *Sunny extremism dataset collects texts from sites hosting known terrorist materials, supplemented with a set of Tumblr posts and Dabiq, Rumiyah, and Inspire issues.*
- *The White Nationalism dataset collects posts from the neonazi forums Stormfront and Vanguard News Network and an existing hate speech dataset.*
- *The Antifascist dataset includes messages from several antifascist networks (e.g., London antifascits, Nomattimen, Antifascist Network, and RevLeft).*
- *The Sovereign Citizen Extremist Dataset collects texts from known forums and websites.*
- *The Benign Dataset from non-radical news articles and Wikipedia.*

Regarding propaganda, there are several available datasets at the document level (TSHP-17 (Rashkin et al., 2017) and Proppy (Barrón-Cedeno et al., 2019)) and fine-grained level (PTC (Martino et al., 2019)). For example, Rashkin et al. (2017) have released the dataset TSHP-17, which contains a balanced dataset of messages classified into four classes:

- *Satire: real news that the reader will not take seriously.*
- *Hoax: aim at convincing readers of a paranoia-fueled story*.
- *Propaganda: aim to mislead readers so that they believe a narrative.*
- *Trusted: real news.*

The authors also provide a lexicon of language found in unreliable articles, classified as hoax and propaganda.

The Proppy corpus Barrón-Cedeno et al. (2019) is a dataset that classifies news as propaganda and non-propaganda. It consists of 52k articles from 100+ news outlets labeled using distant supervision as propaganda or non-propaganda according to the labels of the sources provided by Media Bias/Fact Check (MBFC).

The Proppy corpus has been translated into the Urdu Language by Kausar et al. (2020). Their ProSoul dataset aims at detecting propaganda in the news in the Urdu language. To this end, they have extended the Proppy corpus with 6k non-propaganda and 5k propaganda news.

The PTC corpus (Martino et al., 2019) was created due to the NLP4IF-2019 Shared Task on Fine-Grained Propaganda Detection, held at the conference SemEval. The dataset provides sentence-level annotation (for detecting propaganda sentences) and fine-grained propaganda annotation. The dataset includes 18 types of propaganda techniques: loaded language, name-calling or labeling, repetition, exaggeration or minimization, doubt, appeal to fear/prejudice, flag-waving, causal oversimplification, slogans, appeal to authority, black-and-white fallacy or dictatorship, through-terminating cliché, whataboutism, reductio ad Hitlerum, red herring, bandwagon, obfuscation or intentional vagueness or confusion, and straw man. In addition, they use the media fact check website MBFC[1] to obtain gold labels about propaganda news.

Tundis et al. (2020) use the PTC corpus for detecting mixed-code text, which is a technique used by terrorists to hide their messages. This technique uses special characters to write words to resemble the original intended word. In this work, they have modified a subset of the PTC corpus by replacing one word with a generated art form word in every sentence.

In the context of fake news, Gruppi et al. (2021) have released the NELAGT-2020 Dataset consisting of 1.8M news articles, including embedded tweets from 519 sources annotated with source-level ground truth labels from MBFC for covering multiple dimensions of veracity. In particular, the dataset includes the MBFC factuality score on a scale from 0 to 5 (low to high credibility) and the MBFC Conspiracy/Pseudoscience and questionable sources score, whose value represents low credibility if a source belongs to these categories.

H. Johnston & Weiss (n.d.) classify Sunni propaganda based on a dataset of propaganda (Dabiq, Rumiyah, and Inspire) and a dataset of benign texts (news articles). Unfortunately, the dataset is not available.

As previously mentioned, detecting persuasion techniques in multimedia sources is a new area of research. A new SemEval shared task was introduced in 2021 by Dimitrov et al. (2021). The task is divided into three subtasks: (i) given the textual content of a meme, identify the persuasion technique; (ii) given the textual content of a meme, identify the persuasion techniques used as well as the text spans where they are used; and (iii) given both the text and the image of a meme, identify the persuasion (i.e., propaganda) technique used. The first two subtasks consider 20 persuasion techniques, and the last one, 22, extend the classification proposed by Martino et al. (2019). The task provides a dataset that consists of 950 memes extracted from 26 public Facebook groups. Another available dataset related to memes is HarMeme (Pramanick et al., 2021), which contains 3,544 memes related to COVID-19 collected in Google images, and annotated with the intensity of the label (harmless, partially harmful, and very harmful) and target (individual, organization, community, and society).

Chang & Lin (2021) follow an interesting approach to detect propaganda in the social network Reddit during the COVID-19 pandemic. First, they developed a (non-available) dataset of neutral and pro-china messages in several Reddit threads and developed a propaganda classifier. Then they research cross-platform issues and use this classifier in the social network Twitter. In addition, they complement this analysis by identifying bot accounts using the service Botometer [2] and spreading propaganda. Some of the insights of this study are that pro-china accounts have a higher rate of publishing, tend to publish more on political issues, and exhibit a more negative attitude.

Lastly, Kayode-Adedeji et al. (2019) provide a dataset of 150 mass media YouTube videos on Al-Shahab, Boko Haram, and Islamic State (IS) terrorist groups from 2014 to 2016, classified into 13 subtopics (e.g., advocacy for terrorism, recruitment by terrorists, condemnation of terrorist attacks, ..). This dataset can provide insights into discussions about terrorist groups online.

A different line of research is the generation of counter-narratives. Chung et al. (2019) propose using natural language generation techniques for generating pairs of hate speech and counter-narrative messages.

[1] Available at http://mediabiasfactcheck.com/

[2] Available at https://botometer.osome.iu.edu/

Dataset	Source	Description	Reference
-	Survey	Survey about online extremism detection.	Gaikwad et al. (2021a)
-	Survey	Survey about computational methods for propaganda detection.	Martino et al. (2019)
-	Survey	Survey about multimodal disinformation	Alam et al. (2021)

		detection.	
Datasets Fake News	News	Dataset repository	D'Ulizia et al. (2921)
Dark web recruitment	Forums	730 messages from five dark web forums annotated manually as recruitment and not recruitment messages. Messages are available but not the annotations.	Saini & Bansal (2021)
Recruitment dataset	Forums	This work analyzed 192 messages from the dark web that are annotated as recruitment and not recruitment. The original dataset is the western jihadist website Ansar AlJihad Network.	Scanlon & Gerber (2014)
Extremist dataset	Forums	Non-available datasets from different extremisms: sunny, white nationalism, antifascist and sovereign citizens	A. Johnston & Marku (2020)
TSHP-17	News	Balanced dataset 22580 articles distributed in four classes (trusted, satire, hoax, and propaganda).	Rashkin et al. (2017)
Proppy Corpus	News	52k articles from 100+ news outlets labeled as propaganda or non-propaganda.	Barrón-Centeno et al. (2019)
PTC Propaganda	News	451 articles classified at a fine-grained level according to 18 propaganda techniques	Martino et al. (2019)
ProSoul Dataset	News	Dataset for detecting propaganda in the Urdu language is based on the Proppy corpus.	Kausar et al. (2020)
NELAGT2020	News	Dataset consisting of 1.8M news articles from 519 sources annotated with source-level ground truth labels from MBFC for covering multiple dimensions of veracity.	Gruppi et al. (2021)
Sunni Propaganda	News	Propaganda classification is based on a balanced dataset of propaganda (Dabiq, Rumiyah and Inspire) and benign texts (news articles).	A. H. Johnston & Weiss (n.d.)
SemEval2021 memes	Facebook	950 memes annotated with 22 propaganda techniques.	Dimitrov et al. (2021)
Harmeme	Images	3,544 harmful memes related to COVID-19 collected mainly in Google Images	Pramanick et al. (2021)
YouTube videos	YouTube	150 mass media YouTube videos on Al-Shahab, Boko Haram, and IS terrorist groups from 2014 to 2016 to ascertain	Kayode-Adedeji et al. (2019)

		the kind of discussions about terrorist groups online.	

Table 1: Related work

3. Participation Propaganda and Recruitment Monitoring

This section describes the software architecture of the developed software artifacts developed. First, Sect. 3.1 presents how the software architecture developed in T4.2 for monitoring extremism has been extended for detecting propaganda. In particular, we have included a propaganda classifier module within the data analysis pipeline, which is detailed in Sect. 4. Then, Sect. 3.2 introduces two main changes. First, the propaganda classification can be observed in a widget on the main screen. In addition, we have added two specific dashboards for two specific case studies: the analysis of the religious extremist group “Islamogram” on Reddit and the Italian far-right movement “Mattonisti”. Finally, a new tool, the Participation Chrome Plugin, is described in Sect. 3.3. This plugin takes advantage of the service-based architecture and provides an easy-to-use tool so that end users can analyze the web pages they are browsing.

3.1 Architecture

As described in D4.2, an intelligent engine has been developed as part of the project. This engine allows us to perform thorough monitoring of a large number of metrics that profile language and social media use, among other variables. During task T4.3, the functionalities of the intelligent engine have mainly been expanded to include a broader analysis and two new data sources. The scraping and analysis module have been expanded more concretely, as Figure 1 shows.

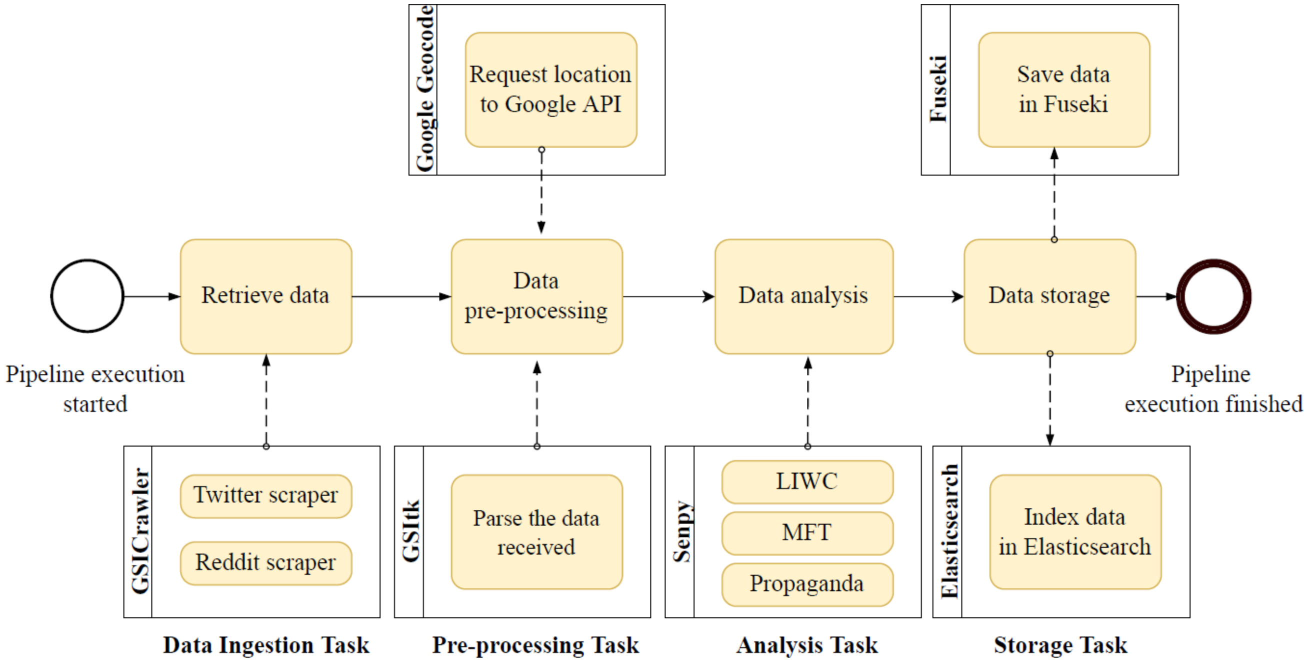


Figure 1: Architecture representation of the proposed model

Scraper module. In task T4.3, a new data source has been added: the Reddit [3] social network. Reddit has been previously studied as an interesting data source for radicalization (Grover & Mark, 2019) and propaganda Balalau & Horincar (2021); Richardson (2020). This site is organized in so-called *subreddits*, communities where a specific topic is usually shared and discussed.

We have detected a set of potentially interesting communities to monitor with the intelligent engine, such *as r/HalalJihadis* [4], *r/SalafisUnveiled* [5], and *r/islamogram* [6]. While Twitter is a more studied platform, Reddit is still a relatively new study objective in the literature. Therefore, we believe it is of great interest to study the Reddit data source, as it offers a new perspective on a different set of users, adding new knowledge.

As part of the efforts of T4.3, different data sources have been considered to be added to the intelligent engine. For example, we have developed a prototype for scraping Discord [7]. Discord has been considered an application used to spread extremism (O'Connor, 2021a). This platform is organized around servers, where one or more topics can be discussed. Unfortunately, most of these servers require a previous invitation, especially those that may show potential extreme discourses. Following these efforts, we have also analyzed the plausibility of scraping from Twitch [8], a popular game streaming platform that has also been studied for extreme language use (O'Connor, 2021b). In this case, the nature of this platform does not allow for its scraping using our system since Twitch can be scraped in a streaming fashion solely.

Analysis module. Following the enhancements made in T4.3, the analysis module has been expanded. We have introduced several machine learning models, including a deep learning approach, to analyze the text captured by our platform and predict whether the mentioned text is propaganda. As part of the development of this module, we propose several combinations of novel machine learning models to address the task of predicting propaganda in text. Also, an extensive experimental evaluation has been performed to assess the performance of said models. Section 4 fully describes this work.

Additionally, several other language analyses have been implemented in the intelligent engine. These analyses give further insight into the language of the data since they study extreme and offensive characteristics of language. All these analysis modules have been implemented as part of the Senpy service (Sánchez-Rada et al., 2020). More concretely, the added analyses are: (i) extreme sentiment through ExtremeSentiLex (Moves, 2022), (ii) grievance-fueled violence (van der Vegt et al., 2021), and (iii) hate-speech analysis using Hurtlex (Bassignana et al., 2018) and the Davidson lexicon (Davidson et al., 2017).

Orchestration. The intelligent engine scrapes the defined data sources periodically and said period could be configured. In order to ease the expansion, configuration and use of the intelligent engine, the orchestration implementation has been shifted to use Airflow (Apache, 2022). This framework allows developers and users to have greater flexibility.

[3] https://reddit.com

[4] https://www.reddit.com/r/HalalJihadis

[5] https://www.reddit.com/r/SalafisUnveiled

[6] https://www.reddit.com/r/islamogram/

[7] https://discord.com/

[8] https://www.twitch.tv/

3.2. Dashboard

The dashboard developed during T4.3 extends the previous dashboard (see D4.2). The new version includes a richer visualization with additional data representations that offer a more detailed data view. Besides, the entire data processing has been improved, which lowers the loading time and thus the usability of the dashboard. Furthermore, the dashboard shows relevant information for several use cases (see Sect. 5). To navigate these use cases, a new heading of the dashboard has been developed, as seen in Figure 2. As seen, users can visit different use cases efficiently.

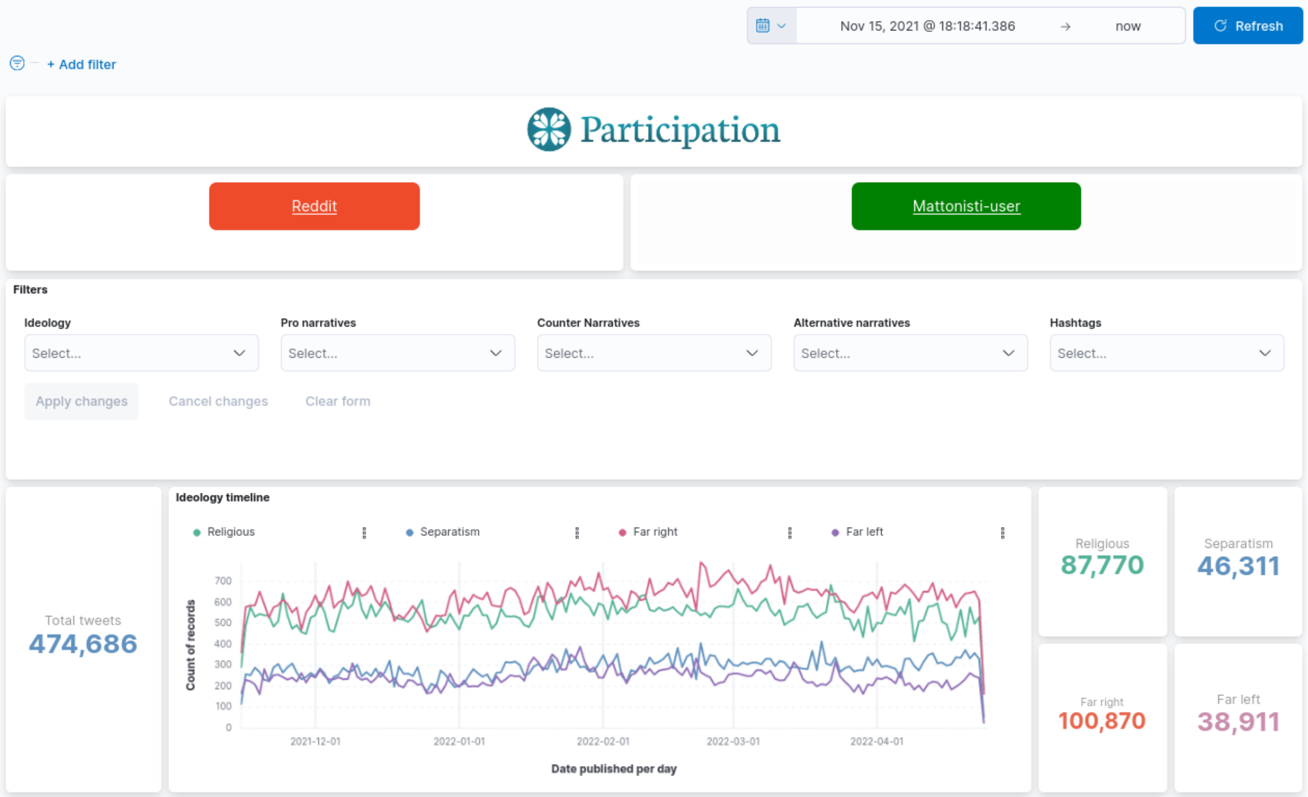


Figure 2: Heading of the dashboard

Two concepts have been introduced to the dashboard to offer a broader view of the data. On the one hand, the dashboard shows the result of computing popular terms, which are those terms that have a high frequency of appearance in the captured data. Figure 3 shows a graph that shows the evolution of popular terms in a selected timeframe.

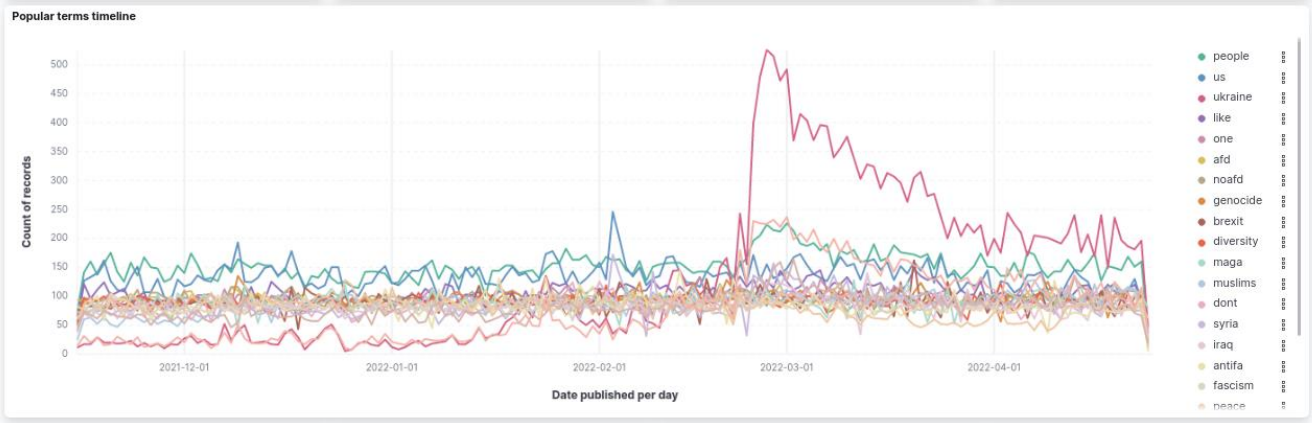


Figure 3: Popular terms timeline graph in the dashboard

On the other hand, the dashboard also contains relevant terms. This last type refers to the terms that are not necessarily frequent in a particular set of documents but characteristic of said set. In this way, users can quickly identify both frequent and characteristic terms from the data. Figure 4 illustrates an example of the visualization of the relevant term.

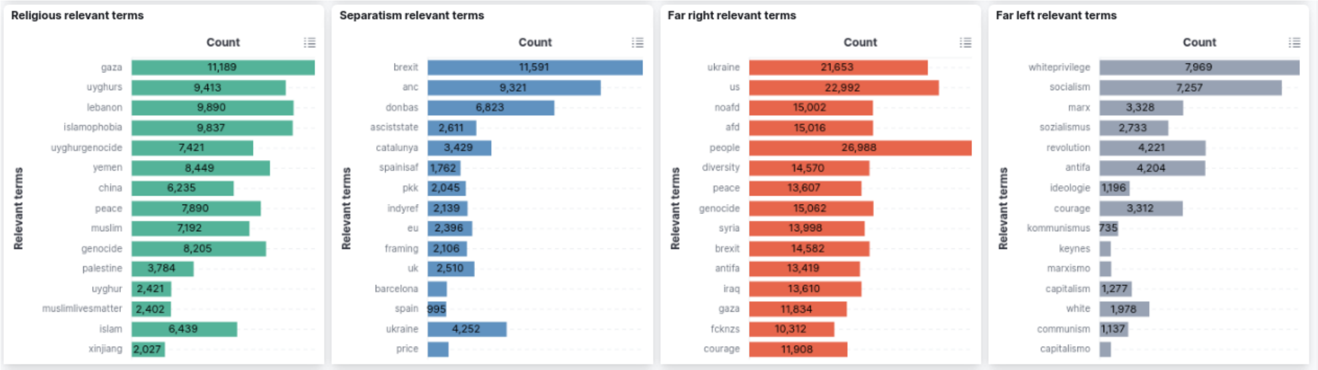


Figure 4: Relevant terms in the dashboard

The digits refer to the number of appearances in the data. As explained, this task introduces the classification of propaganda through machine learning techniques. In addition, a new graph has been added to allow users visualize and filter the data using such information. Figure 5 shows the mentioned visualization.

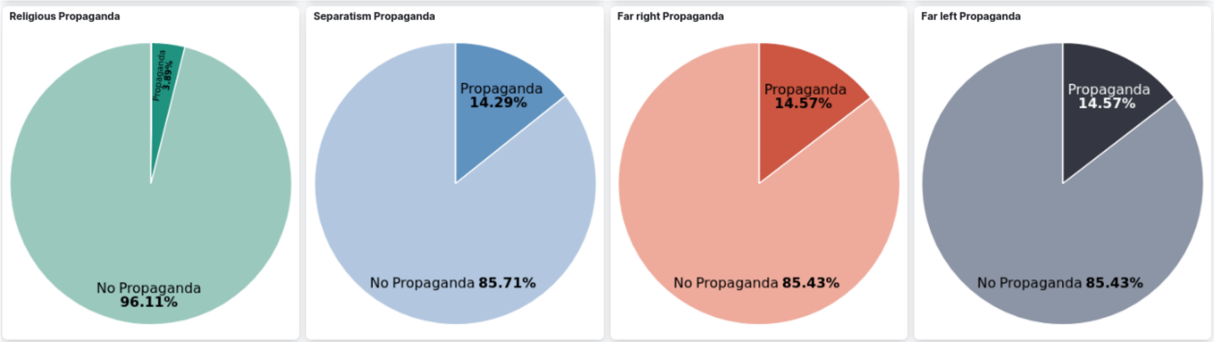
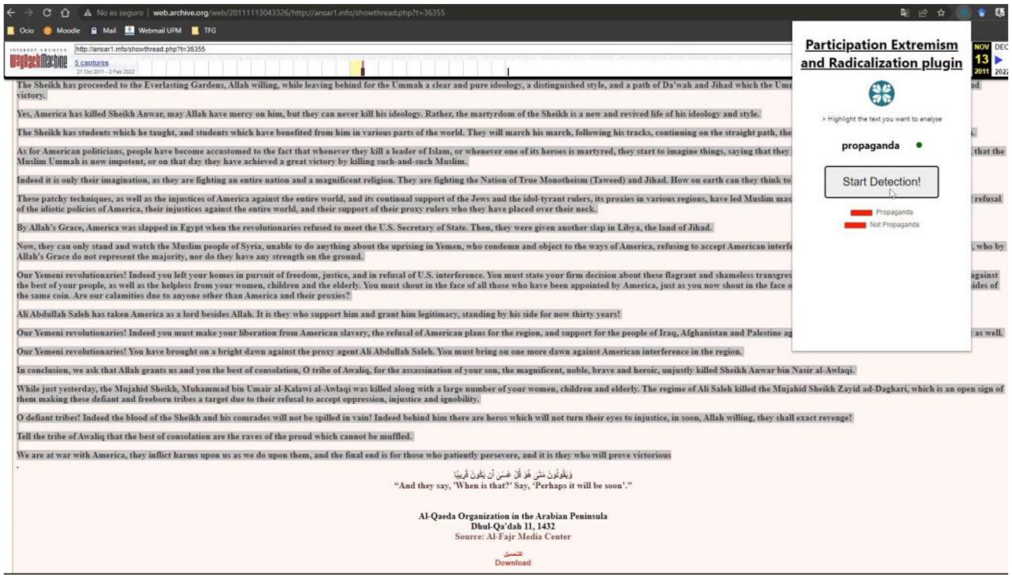


Figure 5: Propaganda analysis results in the dashboard

3.3. Participation Chrome Plugin

In addition to the dashboard described previously, a Chrome Plugin has been developed to enable Participation users to use easy analysis capabilities while browsing a web page. Users can also analyze pages such as emails or their Facebook wall, which could not be analyzed using the dashboard because of security and ethical concerns.

Other analyses can be observed in the plugin documentation available on the Participation webpage.


No es seguro | web.archive.org/web/20111113043326/http://ansar1.info/showthread.php?t=36355
Ocio
Moodle
Mail
Webmail UPM
TFG
INTERNET ARCHIVE
WayBackMachine
http://ansar1.info/showthread.php?t=36355
5 captures
21 Oct 2011 - 2 Feb 2022
NOV
13
2011
DEC
2022
Participation Extremism and Radicalization plugin
> Highlight the text you want to analyse
propaganda
Start Detection!
Propaganda
Not Propaganda

The Sheikh has proceeded to the Everlasting Gardens, Allah willing, while leaving behind for the Ummah a clear and pure ideology, a distinguished style, and a path of Da'wah and Jihad which the Um ... nd victory.

Yes, America has killed Sheikh Anwar, may Allah have mercy on him, but they can never kill his ideology. Rather, the martyrdom of the Sheikh is a new and revived life of his ideology and style.

The Sheikh has students which he taught, and students which have benefited from him in various parts of the world. They will march his march, following his tracks, continuing on the straight path, the

As for American politicians, people have become accustomed to the fact that whenever they kill a leader of Islam, or whenever one of its heroes is martyred, they start to imagine things, saying that they ... that the Muslim Ummah is now impotent, or on that day they have achieved a great victory by killing such-and-such Muslim.

Indeed it is only their imagination, as they are fighting an entire nation and a magnificent religion. They are fighting the Nation of True Monotheism (Taweed) and Jihad. How on earth can they think to

These patchy techniques, as well as the injustices of America against the entire world, and its continual support of the Jews and the idol-tyrant rulers, its proxies in various regions, have led Muslim mas ... refusal of the idiotic policies of America, their injustices against the entire world, and their support of their proxy rulers who they have placed over their neck.

By Allah's Grace, America was slapped in Egypt when the revolutionaries refused to meet the U.S. Secretary of State. Then, they were given another slap in Libya, the land of Jihad.

Now, they can only stand and watch the Muslim people of Syria, unable to do anything about the uprising in Yemen, who condemn and object to the ways of America, refusing to accept American interf ... , who by Allah's Grace do not represent the majority, nor do they have any strength on the ground.

Our Yemeni revolutionaries! Indeed you left your homes in pursuit of freedom, justice, and in refusal of U.S. interference. You must state your firm decision about these flagrant and shameless transgres ... against the best of your people, as well as the helpless from your women, children and the elderly. You must shout in the face of all those who have been appointed by America, just as you now shout in the face o ... ides of the same coin. Are our calamities due to anyone other than America and their proxies?

Ali Abdullah Saleh has taken America as a lord besides Allah. It is they who support him and grant him legitimacy, standing by his side for now thirty years!

Our Yemeni revolutionaries! Indeed you must make your liberation from American slavery, the refusal of American plans for the region, and support for the people of Iraq, Afghanistan and Palestine ag ... as well.

Our Yemeni revolutionaries! You have brought on a bright dawn against the proxy agent Ali Abdullah Saleh. You must bring on one more dawn against American interference in the region.

In conclusion, we ask that Allah grants us and you the best of consolation, O tribe of Awaliq, for the assassination of your son, the magnificent, noble, brave and heroic, unjustly killed Sheikh Anwar bin Nasir al-Awlaqi.

While just yesterday, the Mujahid Sheikh, Muhammad bin Umair al-Kalawi al-Awlaqi was killed along with a large number of your women, children and elderly. The regime of Ali Saleh killed the Mujahid Sheikh Zayid ad-Daghari, which is an open sign of them making these defiant and freeborn tribes a target due to their refusal to accept oppression, injustice and ignobility.

O defiant tribes! Indeed the blood of the Sheikh and his comrades will not be spilled in vain! Indeed behind him there are heros which will not turn their eyes to injustice, in soon, Allah willing, they shall exact revenge!

Tell the tribe of Awaliq that the best of consolation are the raves of the proud which cannot be muffled.

We are at war with America, they inflict harms upon us as we do upon them, and the final end is for those who patiently persevere, and it is they who will prove victorious

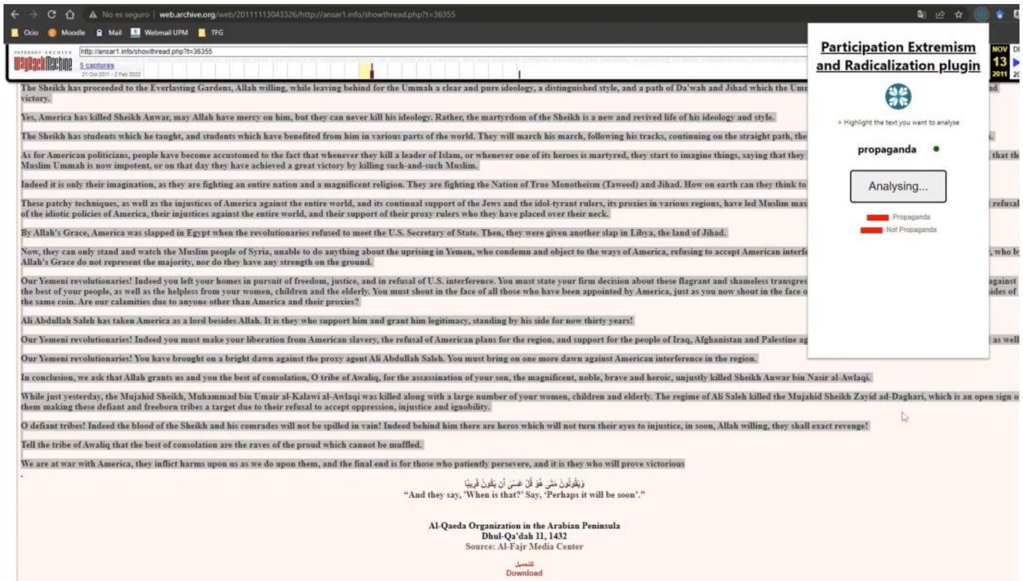
.

وَيَقُولُونَ مَتَىٰ هُوَ ۖ قُلْ عَسَىٰ أَن يَكُونَ قَرِيبًا
"And they say, 'When is that?' Say, 'Perhaps it will be soon'."

Al-Qaeda Organization in the Arabian Peninsula
Dhul-Qa'dah 11, 1432
Source: Al-Fajr Media Center

للتحميل
Download

Figure 6: Launching the Participation Chrome plugin



web.archive.org/web/20111113043326/http://ansar1.info/showthread.php?t=36355

http://ansar1.info/showthread.php?t=36355

The Sheikh has proceeded to the Everlasting Gardens, Allah willing, while leaving behind for the Ummah a clear and pure ideology, a distinguished style, and a path of Da'wah and Jihad which the Umr nd victory.

Yes, America has killed Sheikh Anwar, may Allah have mercy on him, but they can never kill his ideology. Rather, the martyrdom of the Sheikh is a new and revived life of his ideology and style.

The Sheikh has students which he taught, and students which have benefited from him in various parts of the world. They will march his march, following his tracks, continuing on the straight path, the

As for American politicians, people have become accustomed to the fact that whenever they kill a leader of Islam, or whenever one of its heroes is martyred, they start to imagine things, saying that they that th Muslim Ummah is now impotent, or on that day they have achieved a great victory by killing such-and-such Muslim.

Indeed it is only their imagination, as they are fighting an entire nation and a magnificent religion. They are fighting the Nation of True Monotheism (Taweed) and Jihad. How on earth can they think to

These patchy techniques, as well as the injustices of America against the entire world, and its continual support of the Jews and the idol-tyrant rulers, its proxies in various regions, have led Muslim mas refusal of the idiotic policies of America, their injustices against the entire world, and their support of their proxy rulers who they have placed over their neck.

By Allah's Grace, America was slapped in Egypt when the revolutionaries refused to meet the U.S. Secretary of State. Then, they were given another slap in Libya, the land of Jihad.

Now, they can only stand and watch the Muslim people of Syria, unable to do anything about the uprising in Yemen, who condemn and object to the ways of America, refusing to accept American interfe , who b Allah's Grace do not represent the majority, nor do they have any strength on the ground.

Our Yemeni revolutionaries! Indeed you left your homes in pursuit of freedom, justice, and in refusal of U.S. interference. You must state your firm decision about these flagrant and shameless transgres against the best of your people, as well as the helpless from your women, children and the elderly. You must shout in the face of all those who have been appointed by America, just as you now shout in the face o sides of the same coin. Are our calamities due to anyone other than America and their proxies?

Ali Abdullah Saleh has taken America as a lord besides Allah. It is they who support him and grant him legitimacy, standing by his side for now thirty years!

Our Yemeni revolutionaries! Indeed you must make your liberation from American slavery, the refusal of American plans for the region, and support for the people of Iraq, Afghanistan and Palestine ag as well

Our Yemeni revolutionaries! You have brought on a bright dawn against the proxy agent Ali Abdullah Saleh. You must bring on one more dawn against American interference in the region.

In conclusion, we ask that Allah grants us and you the best of consolation, O tribe of Awaliq, for the assassination of your son, the magnificent, noble, brave and heroic, unjustly killed Sheikh Anwar bin Nasir al-Awlaqi.

While just yesterday, the Mujahid Sheikh, Muhammad bin Umair al-Kalawi al-Awlaqi was killed along with a large number of your women, children and elderly. The regime of Ali Saleh killed the Mujahid Sheikh Zayid ad-Daghari, which is an open sign o them making these defiant and freeborn tribes a target due to their refusal to accept oppression, injustice and ignobility.

O defiant tribes! Indeed the blood of the Sheikh and his comrades will not be spilled in vain! Indeed behind him there are heros which will not turn their eyes to injustice, in soon, Allah willing, they shall exact revenge!

Tell the tribe of Awaliq that the best of consolation are the raves of the proud which cannot be muffled.

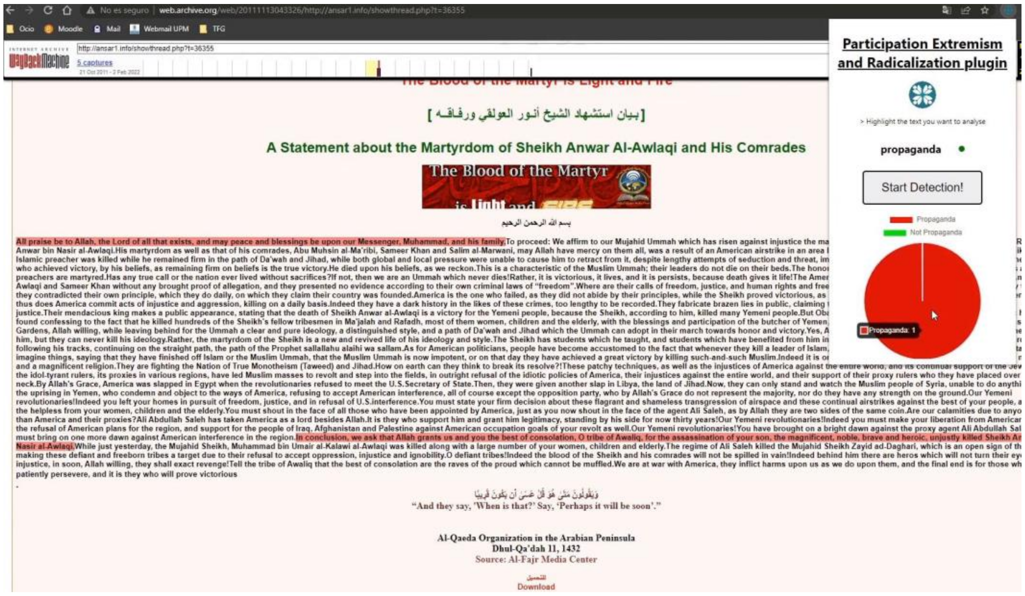
We are at war with America, they inflict harms upon us as we do upon them, and the final end is for those who patiently persevere, and it is they who will prove victorious

وَيَقُولُونَ مَتَىٰ هُوَ قُلْ عَسَىٰ أَن يَكُونَ قَرِيبًا
“And they say, 'When is that?' Say, ‘Perhaps it will be soon’.”

Al-Qaeda Organization in the Arabian Peninsula
Dhul-Qa'dah 11, 1432
Source: Al-Fajr Media Center

للتحميل
Download

Figure 7: Starting the analysis



web.archive.org/web/20111113043326/http://ansar1.info/showthread.php?t=36355

http://ansar1.info/showthread.php?t=36355

The Blood of the Martyr is Light and Fire

[بيان استشهاد الشيخ أنور العولقي ورفاقه]

A Statement about the Martyrdom of Sheikh Anwar Al-Awlaqi and His Comrades

بسم الله الرحمن الرحيم

All praise be to Allah, the Lord of all that exists, and may peace and blessings be upon our Messenger, Muhammad, and his family.To proceed: We affirm to our Mujahid Ummah which has risen against injustice the ma R
Anwar bin Nasir al-Awlaqi.His martyrdom as well as that of his comrades, Abu Muhsin al-Ma'ribi, Sameer Khan and Salim al-Marwani, may Allah have mercy on them all, was a result of an American airstrike in an area l
Islamic preacher was killed while he remained firm in the path of Da'wah and Jihad, while both global and local pressure were unable to cause him to retract from it, despite lengthy attempts of seduction and threat, im
who achieved victory, by his beliefs, as remaining firm on beliefs is the true victory.He died upon his beliefs, as we reckon.This is a characteristic of the Muslim Ummah; their leaders do not die on their beds.The honor
preachers are martyred.Has any true call or the nation ever lived without sacrifices?If not, then we are an Ummah which never dies!Rather, it is victorious, it lives, and it is persists, because death gives it life!The Amer
Awlaqi and Sameer Khan without any brought proof of allegation, and they presented no evidence according to their own criminal laws of “freedom”.Where are their calls of freedom, justice, and human rights and free
they contradicted their own principle, which they do daily, on which they claim their country was founded.America is the one who failed, as they did not abide by their principles, while the Sheikh proved victorious, as
thus does America commit acts of injustice and aggression, killing on a daily basis.Indeed they have a dark history in the likes of these crimes, too lengthy to be recorded.They fabricate brazen lies in public, claiming
justice.Their mendacious king makes a public appearance, stating that the death of Sheikh Anwar al-Awlaqi is a victory for the Yemeni people, because the Sheikh, according to him, killed many Yemeni people.But Oba
found confessing to the fact that he killed hundreds of the Sheikh's fellow tribesmen in Ma'jalah and Rafadh, most of them women, children and the elderly, with the blessings and participation of the butcher of Yemen,
Gardens, Allah willing, while leaving behind for the Ummah a clear and pure ideology, a distinguished style, and a path of Da'wah and Jihad which the Ummah can adopt in their march towards honor and victory.Yes, A
him, but they can never kill his ideology.Rather, the martyrdom of the Sheikh is a new and revived life of his ideology and style.The Sheikh has students which he taught, and students which have benefited from him in
following his tracks, continuing on the straight path, the path of the Prophet sallallahu alaihi wa sallam.As for American politicians, people have become accustomed to the fact that whenever they kill a leader of Islam,
imagine things, saying that they have finished off Islam or the Muslim Ummah, that the Muslim Ummah is now impotent, or on that day they have achieved a great victory by killing such-and-such Muslim.Indeed it is or
and a magnificent religion.They are fighting the Nation of True Monotheism (Taweed) and Jihad.How on earth can they think to break its resolve?!These patchy techniques, as well as the injustices of America against the entire world, and its continual support of the Jew
the idol-tyrant rulers, its proxies in various regions, have led Muslim masses to revolt and step into the fields, in outright refusal of the idiotic policies of America, their injustices against the entire world, and their support of their proxy rulers who they have placed over
neck.By Allah's Grace, America was slapped in Egypt when the revolutionaries refused to meet the U.S.Secretary of State.Then, they were given another slap in Libya, the land of Jihad.Now, they can only stand and watch the Muslim people of Syria, unable to do anythi
the uprising in Yemen, who condemn and object to the ways of America, refusing to accept American interference, all of course except the opposition party, who by Allah's Grace do not represent the majority, nor do they have any strength on the ground.Our Yemeni
revolutionaries!Indeed you left your homes in pursuit of freedom, justice, and in refusal of U.S.interference.You must state your firm decision about these flagrant and shameless transgression of airspace and these continual airstrikes against the best of your people, a
the helpless from your women, children and the elderly.You must shout in the face of all those who have been appointed by America, just as you now shout in the face of the agent Ali Saleh, as by Allah they are two sides of the same coin.Are our calamities due to anyo
than America and their proxies?Ali Abdullah Saleh has taken America as a lord besides Allah.It is they who support him and grant him legitimacy, standing by his side for now thirty years!Our Yemeni revolutionaries!Indeed you must make your liberation from Americar
the refusal of American plans for the region, and support for the people of Iraq, Afghanistan and Palestine against American occupation goals of your revolt as well.Our Yemeni revolutionaries!You have brought on a bright dawn against the proxy agent Ali Abdullah Sal
must bring on one more dawn against American interference in the region.In conclusion, we ask that Allah grants us and you the best of consolation, O tribe of Awaliq, for the assassination of your son, the magnificent, noble, brave and heroic, unjustly killed Sheikh Ar
Nasir al-Awlaqi.While just yesterday, the Mujahid Sheikh, Muhammad bin Umair al-Kalawi al-Awlaqi was killed along with a large number of your women, children and elderly.The regime of Ali Saleh killed the Mujahid Sheikh Zayid ad-Daghari, which is an open sign of th
making these defiant and freeborn tribes a target due to their refusal to accept oppression, injustice and ignobility.O defiant tribes!Indeed the blood of the Sheikh and his comrades will not be spilled in vain!Indeed behind him there are heros which will not turn their ey
injustice, in soon, Allah willing, they shall exact revenge!Tell the tribe of Awaliq that the best of consolation are the raves of the proud which cannot be muffled.We are at war with America, they inflict harms upon us as we do upon them, and the final end is for those wh
patiently persevere, and it is they who will prove victorious

وَيَقُولُونَ مَتَىٰ هُوَ قُلْ عَسَىٰ أَن يَكُونَ قَرِيبًا
“And they say, 'When is that?' Say, ‘Perhaps it will be soon’.”

Al-Qaeda Organization in the Arabian Peninsula
Dhul-Qa'dah 11, 1432
Source: Al-Fajr Media Center

للتحميل
Download

Figure 8: Showing results of the analysis.

4. Machine learning-based Propaganda classifier

Considering tasks T4.1 and T4.2, an in-depth analysis of radical propaganda has been made as part of T4.3. In this sense, inspired by the literature (Sect. 2) and tasks T4.1 and T4.2, we have considered various signals that aid to model radical processes. An intelligent learning system has been developed to detect radical propaganda from the text. This system leverages the information extracted from text, contextualizing it with several additional signals: (i) emotion and (ii) moral values signals, (iii) similarity-based features that exploit a word embedding model, (iv) powerful word combinations through relevant bigram detection, and (v) different feature combinations that aim to improve the system's overall performance. In addition to this, T4.3 also explores the effectiveness of a deep learning model: Bidirectional Encoder Representations from Transformers (BERT).

4.1. Ensemble and distributional learning model

We propose a machine learning model that leverages several knowledge sources to improve its performance. As seen in Figure 9, this model comprises four different information sources that offer a complete view of the analyzed text. The input text is processed by the four feature extractors, which are then combined into a unified vector representation and fed to a machine learning classifier. Finally, this learner outputs a prediction based on combining the information gathered by the produced features.

A full description of the developed models for the interested reader can be found in Araque & Iglesias (n.d.). Following, we describe the primary information sources and how they are obtained.

Emotion signals. Previous works discuss how emotions and moral values can be exploited to spread propaganda and how this is linked to radicalization processes Da San Martino et al. (2021); Decety et al. (2018). Furthermore, previous works leverage emotional information for detecting radicalization Araque & Iglesias (2020), but such approaches have not been extensively studied in the propaganda domain. We model emotional presence with the EmoFeat (Emotion Features) algorithm Araque & Iglesias (2020). EmoFeat uses an emotion lexicon to extract emotion-driven features later fed to a machine learning model. In this way, we propose using a lexicon-based representation that uses statistical measures to encode emotional characteristics in text.

Moral signals. We propose using a lexical resource to characterize moral values present in a certain text to assess moral rhetoric. In this way, a specific moral lexicon is composed of several moral foundations (care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and purity/degradation). For each moral foundation, a lexical resource contains a set of words that are not necessarily shared among different moral values. The moral values are modeled using the MoralStrength lexicon Oscar Araque et al. (2020). To compute a unified representation of the moral values, the method extracts the average.

$$f = \frac{1}{\|s\|} \sum_{i=1}^{\|s\|} s_i$$

where s is the moral annotations as they appear in the analyzed text. As described, this method computes unified features from texts that express moral presence and intensity, then is used to train several learning models.

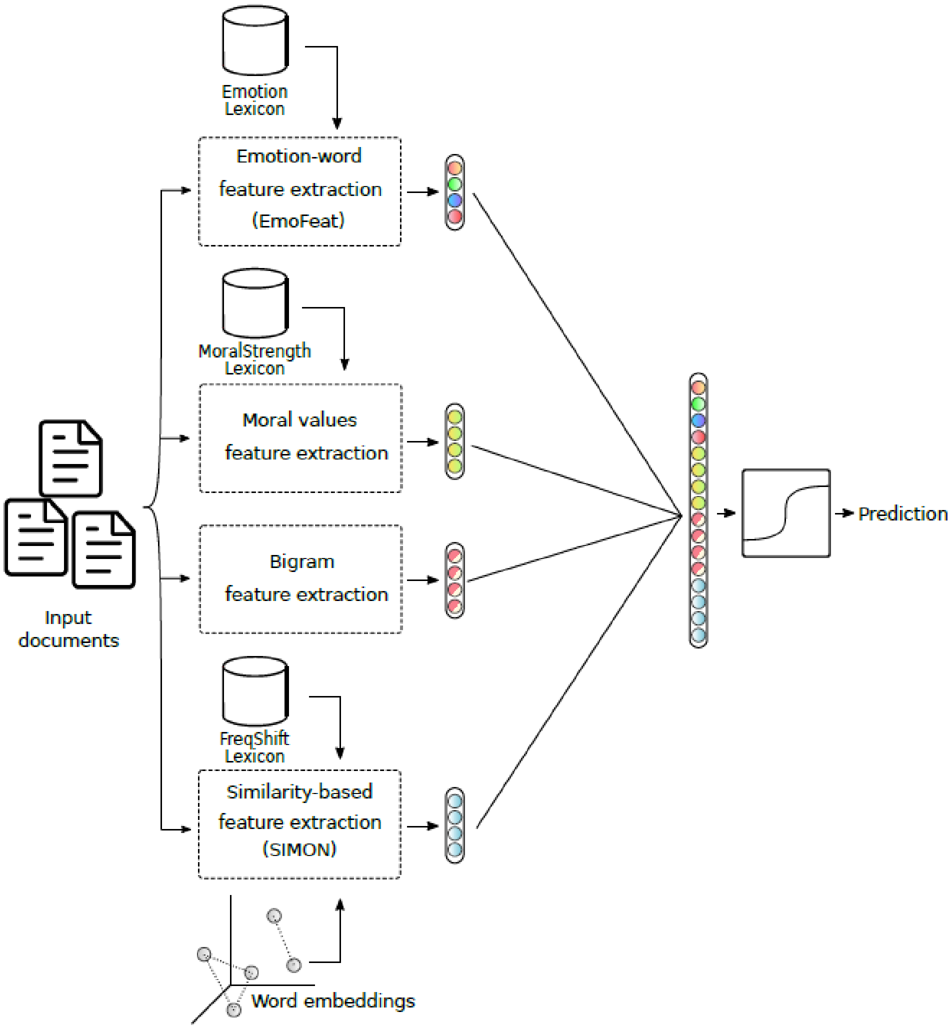

Emotion Lexicon
Emotion-word feature extraction (EmoFeat)
MoralStrength Lexicon
Moral values feature extraction
Input documents
Bigram feature extraction
FreqShift Lexicon
Similarity-based feature extraction (SIMON)
Word embeddings
Prediction

Figure 9: Architecture representation of the proposed model

Distributional semantics and similarity. To include robust text representation, we use the SIMilarity-based sentiment projectiON (SIMON) model Araque et al. (2019). The SIMON method uses a pre-trained word embedding model and a domain-oriented lexicon to compute features for a particular domain. This work adapts the SIMON method in conjunction with propaganda lexicons. As an overview, the SIMON method uses a domain lexicon to measure its similarity against the analyzed text, as done similarly in Araque & Iglesias (2020). This computes a feature vector that measures the similarity projection of the input texts to the chosen lexicon. Using such a method, it is possible to exploit the knowledge contained in word embeddings as well as the domain-oriented information contained in the lexicon. Additionally, the SIMON method does not need large corpora for its training and can be applied in tasks where annotated data is challenging to obtain.

Frequency shift for domain adaptation. Previous methods use FreqSelect, a simple method to implement and can yield reasonably good results Araque & Iglesias (2020). Still, it is interesting that FreqSelect does not consider class-specific word frequencies. The relative frequency of a word may convey information that can be exploited to improve the classification result. That is, the frequency of a word may vary from one class to another, and this knowledge can be used by a learning model.

In order to address this, we propose the use of frequency shifts Gallagher et al. (2021). Using the shift in relative frequencies, we then select the words that have a higher shift value. These shift values can be used to describe differences between domains Gallagher et al. (2021); Dodds et al. (2015). The interpretation of such a measure allows us to obtain information regarding class word frequency, improving the overall system performance. This method is called FreqShift.

Peak pattern detection. Additionally, we consider the effect of adding bigram representations to be used in combination with the SIMON model. It is interesting to study the effect of increasing the coverage for certain pair of tokens on the learning model. These aids the classifier in detecting specific patterns that arise from the analyzed texts. The idea of considering bigrams is common in other Natural Language Processing (NLP) tasks Wang & Manning (2012); Pedersen (2001). This work explores it further when used in combination with the SIMON model.

4.2. Transformers model

Transformers is a deep learning model, mainly used in NLP, which is wholly based on self-attention, i.e., associating different positions of a sequence as a means to compute a representation of it. When used for NLP applications, Transformers are usually considered language models. Furthermore, Transformers divides the input data, weighting the importance of every part in the whole context. As described in Vaswani et al. (2017), it differs from the traditional Recurrent Neural Networks (RNNs) in the fact that Transformers does not demand to process the sequential data in its entry order. Therefore, it can predict the context for words that are placed in every position of the sentence. Precisely, RNNs mainly take into account the symbol positions of both input and output sequences by matching positions into steps in the computation. However, the main obstacle to this sequential method is that it is incompatible with a critical feature when training long sequences: parallelization.

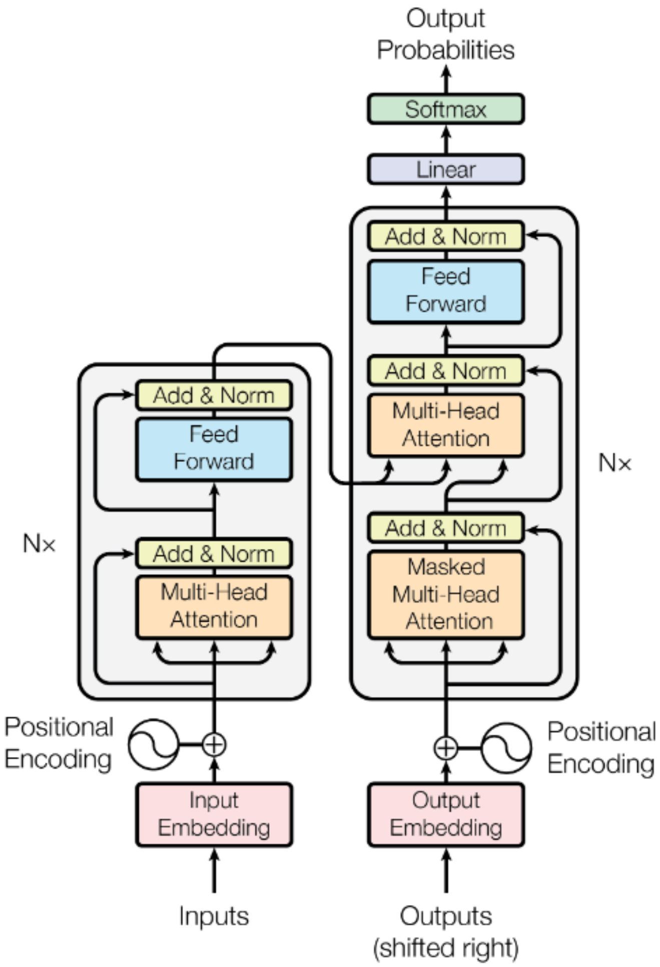


Figure 10: The model architecture of the Transformer (Vaswani et al., 2017).

As mentioned, the transformer model architecture avoids sequential RNNs or convolution, depending just on the relationships which can be formed among every single feature of input and output. As far as it is concerned, it is the first deep learning model to have achieved this objective fully. Similarly to the most valuable neural models, the Transformer is based on an encoder-decoder structure, where the first allocates the input symbols to a continuous sequence while the second generates the output by means of this continuous sequence, taking into account that the model absorbs the previously generated symbols as an input in order to produce the next one. Nevertheless, as seen in Figure 10, the essential characteristic of the Transformer architecture is that it includes stacked self-attention, as has already been explained (in fact, both encoder and decoder are formed by N=6 layers). Specifically, each layer of the encoder is composed of two sub-layers, which correspond to multi-head attention and a positioning mechanism. Furthermore, the decoder adds a third sub-layer which executes multi-head attention over the output of the encoder pile.

4.3. Experimental evaluation results

In order to evaluate the performance of the proposed learning-based models, a thorough evaluation has been made in the task of radical propaganda detection. The proposed learning architecture which is proposed can be implemented with several variations (i.e., different word embedding models can be used). To thoroughly evaluate the quality of the architecture, an experimental evaluation has been performed. With the aim of giving complete performance metrics, the weighted averaged f-score is used as the performance metric evaluated over a train-test splitting of the considered datasets (Sect. 1): NELA-GT, QProp, PTC, and Jacobs. Overall, this evaluation considers the effect of the different features of the proposed system. In this way, we study the performance when using single features and with their combinations. As we are interested in studying the effect of the features on the performance of the system, the evaluation uses a classic algorithm, logistic regression.

Table 2 shows the averaged f-score of all considered models and their combinations and their comparison with the selected transformer model. Attending to the results, it can be seen that combining distributional representations (as those obtained by the SIMON model) with contextualizing information (such as emotion, morals, and peak signals through bigrams) can improve the performance of the learning system. Also, the presented proposed FreqShift method generates a domain-adapted lexicon as part of the training process. Such resources can be used for downstream applications without the need for automatic systems that consume high loads of computation resources Araque & Iglesias (n.d.).

Following the comparison against the BERT model, it can be seen that the selected transformer model can obtain higher performance scores in almost all datasets. Still, when evaluating the QProp dataset, the domain-adapted SIMON model combined with bigrams outperforms the BERT model. In light of this, we argue that the application of this kind of language model in the field of radicalization can enhance the performance of existing systems and, generally, lead practitioners to obtain new state-of-the-art automatic systems.

The general applicability of a transformer model to the problem at hand, although experimentally demonstrated, is still an open challenge. New methods that fine-tune and adapt existing language models to a specific domain (as in the case of radicalization) are needed. This can be understood when attending to the fact that the BERT model has not outperformed in all datasets. Future work would need to obtain an explanation for this behavior using a data-driven approach.

Finally, using the previous evaluation as a guide, the intelligent engine makes use of the Transformers model to predict the existence of propaganda in text. To enhance the robustness of the system and to avoid generalization issues, the intelligent engine considers that a text contains propaganda when at least one of the models trained over the considered datasets outputs a positive classification. In this way, the predictions of the four transformer models are combined.

Logistic Regression	NELA-GT	QProp	PTC	Jacobs
MoralStrength (MS)	30.33	63.44	49.78	73.57
Emotion	39.38	69.60	52.08	79.89
Bigram	58.09	88.00	60.54	83.95
Unigram	59.33	89.96	66.31	81.45
SIMON	57.50	89.69	68.63	91.20
Unigram + MS	56.63 (↓2.70)	90.14 (↑0.18)	66.68 (↑0.37)	72.08 (↓9.37)
Unigram + Emotion	57.16 (↓2.17)	89.65 (↓0.31)	66.40 (↑0.09)	73.96 (↓7.49)
SIMON + MS	59.35 (↑1.85)	89.53 (↓0.16)	63.33 (↓5.30)	90.70 (↓0.50)
Simon + Emotion	58.36 (↑0.86)	89.00 (↓0.69)	60.30 (↓8.33)	90.60 (↓0.60)
SIMON + Bigrams	60.78 (↑3.28)	**90.97 (↑1.28)**	59.07 (↓9.56)	83.08 (↓8.12)
Unigram + MS + Emotion	58.05 (↓1.28)	90.49 (↑0.53)	66.78 (↑0.47)	83.97 (↑2.52)
SIMON + MS + Emotion	59.03 (↑1.53)	89.43 (↓0.26)	67.52 (↓1.11)	88.69 (↓2.31)
SIMON + MS + Emo + Bigram	55.66 (↓1.84)	90.88 (↑1.19)	67.42 (↓1.21)	85.50 (↓5.70)
Transformers (BERT)	**71.67**	83.49	**74.84**	**95.44**

Table 2: Averaged F1-scores for the considered models in all datasets using Logistic Regression as classifier

4.4. Manual annotation and evaluation

This section presents an additional evaluation on the developed machine learning model through a human-driven assessment of the propaganda phenomena. To do so, we have selected the two best models described in the previous section (i.e., the *SIMON + Bigrams* and *Transformers (BERT)* approaches using a Logistic Regression classifier) and evaluated with an additional dataset. This dataset has been generated for this purpose, and has been manually annotated.

Firstly, we have captured a representative set of Twitter messages, ranging from January 2020 to May 2021, with the intention of capturing messages related to the COVID pandemic generated by public figures and institutions. Thus, following the methodology presented by Moral et al. (2023), we have addressed messages written in English and originated in Europe, USA, Russia and China. Originally, we have captured 2,567 messages that after several filters have been reduced to a set of 112 messages. These messages are balanced in their representativeness of nationalities.

Following, to assess the presence of propaganda, we have performed a manual annotation of the described data. To do so, seven distinct annotators, with technical knowledge and experience in Natural Language Processing, have been instructed to the task of propaganda annotation. Annotators are asked to identify and indicate whether the data instances reflect propaganda or no propaganda. After the annotation process, it is interesting to assess the agreement among annotators. This kind of measure offers information regarding the result of the annotations, allowing to discard outlier annotations or even an entire annotator.

Concretely, we have used the Cohen's and Fleiss kappa metrics (Artstein et al., 2008). Cohen's metric is defined as follows.

$$\kappa_C = \frac{p_o - p_e}{1 - p_e}$$

where p_o represents the relative observed agreement among all raters; and p_e conveys the computed probability of chance agreement by using the observed data. The probability of change agreement is estimated by assessing the probabilities of each observer randomly seeing each category. Cohen's kappa is between 0 and 1: if $\kappa_C = 1$ the annotators are in complete agreement, and if $\kappa_C = 0$ the annotators do not have a larger agreement than that given by chance. Following, the Fleiss kappa is defined as:

$$\kappa_F = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P}_e}$$

This factor $1 - \overline{P}_e$ expresses the degree of agreement that can be achieved above chance, while the factor $\overline{P} - \overline{P}_e$ conveys the degree of agreement above chance that has been achieved. If the raters are in complete agreement then the derived metric would be $\kappa_F = 1$. If the agreement among raters is that given by chance, then $\kappa_F \leq 0$. These two metrics quantitatively inform us of both the quality of the annotations and the difficulty of the task itself.

Following these definitions, we have obtained an averaged Cohen's kappa of 0.23, and a Fleiss kappa of 0.22. The detailed metrics can be seen in Table 3. In the case of multiple annotators, it is common to compute the Fleiss kappa, even though in this case both averaged Cohen and Fleiss metrics are very similar. Following the interpretation of the Fleiss kappa described by Von Eye et al. (2014), it can be seen that the annotators have reached a “fair agreement”. Considering the considerable number of annotators and the difficulty of the task, we consider this to be a positive result. Following common practice in the field, we have composed a set of gold labels by aggregating the individual annotations into a single label per document. This allows a machine learning system to be trained on such a set of labels or, as in our case, to be evaluated in said set.

Propaganda annotation								
Cohen’s kappa		**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**
	A1	-	0.32	0.19	0.29	0.36	0.08	0.35
	A2	0.32	-	0.17	0.22	0.37	0.20	0.41
	A3	0.19	0.17	-	0.19	0.17	0.03	0.08
	A4	0.29	0.22	0.19	-	0.19	0.10	0.32
	A5	0.36	0.37	0.17	0.19	-	0.16	0.38
	A6	0.08	0.20	0.03	0.10	0.16	-	0.16
	A7	0.35	0.41	0.08	0.32	0.38	0.16	-
	Average: **0.23**							
Fleiss’ kappa	**0.22**							

Table 3: Detailed Cohen’s and Fleiss kappa among all annotators

Continuing with the evaluation, we have used the obtained dataset (both the captured messages and their obtained labels) to evaluate the described machine learning systems. That is, we have not used the derived dataset for any training, but rather used it as test set of the already trained models. Such a setting of the evaluation aims to assess the generalization capabilities of the described models, that is, the capabilities to detect propaganda in data not seen during training by the models.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
SIMON + Bigrams	71.29	70.01	70.12	70.06
Transformers (BERT)	74.13	72.58	73.87	73.22

Table 4: Scores for the two considered models in the new evaluation data

Finally, to better assess the generalization capabilities of the models we compare to two works that perform propaganda classification in different datasets. Please note that this comparison is not direct, since the literature works do not use our newly generate set. This comparison serves as a reference point to understand the current challenges of propaganda detection task.

Therefore, it is interesting to see that Raza (2021) has achieved a f-score of 75.00 in the NELA-GT dataset, while Da San Martino et al. (2019) have obtained in the same metric 63.23 while testing in the PTC data. Considering that these two evaluations are done on the same domain (i.e., the training and test instances are taken from the same dataset), we observe that the results obtained by the proposed systems in a positive light.

Finally, Table 5 shows some examples of the captured data, the annotations that have been generated as well as the golden standard that has been computed for each data instance. This data sample illustrates the difficulty of the task of propaganda annotation and detection.

Text	A1	A2	A3	A4	A5	A6	A7	Agg.
Back in 2003 US/UK claimed Iraq had WMDs. Later they blamed Syria for chemical weapons. In 2014 the West lied on #MH17. In 2018 it hyped on “highly likely” #Scripal case. Now “almost certain” RU hacking #vaccine hoax. What next?	1	1	1	1	0	0	1	1
More lies from #Maduro, author of the #MaduroCrisis, about #Venezuela’s tragedy. U.S. sanctions NEVER block food or medicine. Shortages in Venezuela result from the regime’s theft of the nation’s wealth. #EstamosUnidosVE	1	0	1	1	1	0	1	1
Donations from Chinese entrepreneur, Jack Ma and his Alibaba Group, will arrive in Ireland today, including 300,000 masks, 30,000 testing kits, 3,000 protective suits. The delivery first touched down in Belgium on Mar.23 along with donations to other EU states. A civil plane loaded with masks, is flying from Jiangsu Province to the	0	0	0	0	1	1	0	0

destination—#Wuhan. It’s really a race against time, second by second. #coronavirus								
China is working to take down freedom all across the world.	0	1	1	1	1	1	1	1
A small group of countries, led by the United States, Germany and the United Kingdom, abused the UN platform, politicized the issues of human rights and provoked confrontation. 153 | #ThirdCommittee #75UNGA	1	1	0	1	1	0	1	1
TIME tells all. In Maldives, responses to tackle Covid19 are challenging but impressive. Time will tell We're with You. China proves through time that resolute actions bring results. Certain country has wasted 2-months window time by mocking, blaming, & instigmatization	1	0	0	1	0	0	1	0
With this song in violin, I wish you a happy #PakistanDay! This song is out of the world! I wish Pakistan could defeat #COVID19 ASAP! #StandWithPakistan! Chin-Pakistan dosti zindabad!	0	0	0	0	0	1	0	0
The Chinese Communist Party is behaving in ways that fundamentally put the American people’s security at risk. The @realDonaldTrump Administration is the first in decades to take this threat seriously.	1	1	1	1	1	0	1	1

Table 5: Examples capture for the manual annotation and evaluation. A1 to A7 are the human annotations, while the “Agg.” column is the aggregated golden label.

4.5. Examples

In this section, several examples, extracted from both the dashboard data (where different social media is included, as it is expounded in the next section) and the datasets which have been explained previously, have been analyzed. Indeed, it must be considered when inspecting Table 6 that the first sample linked to a certain origin corresponds to a non-propagandist phrase, whereas the second is related to propaganda content. The aim of these examples is merely to understand the behavior of the models since, after being trained by means of a considerable amount of data, they predict whether this content is indeed propagandist or not. Therefore, they have been selected because it is considerably straightforward for the models to identify which contain propaganda.

Source	Description	Propaganda
NELA-GT	Paris Saint-Germain president Nasser Al-Khelaifi and former FIFA general secretary Jerome Valcke were this morning charged with related offences by the Swiss Attorney General over the awarding of media rights at various international tournaments.	NO
	After four punch-drunk years of Donald Trump, the weeks since the November presidential election have presented a chance, despite his machinations to overturn the result, to reflect on what might come next for the tens of millions of Americans struggling to get by.	YES
QProp	Eat in Connecticut teamed up with End Hunger Connecticut (EHC) for the first-ever Culinary Corner Pub Crawl, Jan. 20. The crawl featured top Blue Back Square pubs and restaurants.	NO
	The threat by President Trump to fire the Federal Reserve chairman is an example of how our entire federal system has grown into a monstrosity that is overwhelmingly unconstitutional - and almost everyone, not just President Trump, is to blame.	YES
PTC	Kristian Saucier, who served a year in federal prison for taking photos of classified sections of the submarine on which he worked, says he was subject to unequal protection of the law.	NO
	Barack Hussein Obama has planted seeds that will be bearing bitter fruit for years, and probably decades, to come.	YES
Jacobs	An Afghan city mayor has been killed after an explosion struck his vehicle as he returned home from work in the east of the country, police say.	NO
	In the name of God the Merciful Praise be to God Moez Islam and Muslims, and humiliating the infidels and apostates, peace and blessings be upon the Imam of Mujahideen, and the leader of the resplendent cheerful fighting.	YES
Twitter	Streamer who wants to get the attention of his followers. So use your sub badges in stream I will create for You DM me.	NO
	@ApostateProphet You came to USA cause you had something to `sell' -	YES

	your fake hate for Muslims. You thought tht was enough to endear you to the Bible thumpers, huh?! Wrong! The current flavour of the season in USA is not hate for Muslims, it's hate for Jews.	
Discord	Religious or non religious, what are your goals for 2022?	NO
	What does this sub make of such news? His actions included promoting atheism.	YES

Table 6: Examples introduced to the models, grouped by their origin

5. Use cases: study of propaganda language in social media

The developed intelligent engine can capture a large set of data from different data sources. Such data is then analyzed and stored in two different databases and can be easily accessed through an interactive web dashboard. To better assess the utility of this platform and follow motivation from the literature and previous project deliverables, three use cases have been defined. The results of each use case can be explored by its corresponding interactive dashboard. Following, the use cases are described.

5.1. Monitoring Propaganda on Twitter

Twitter is a social network where harsh language is spread Benigni et al. (2017). Also, there are a large number of works in the literature that study Twitter content to model extremist behavior Gaikwad et al. (2021b). In light of this, a use case has been defined to monitor four different ideologies that express three distinct narratives (see D4.2). This use case expands on the one initially developed in T4.2 (described in D4.2).

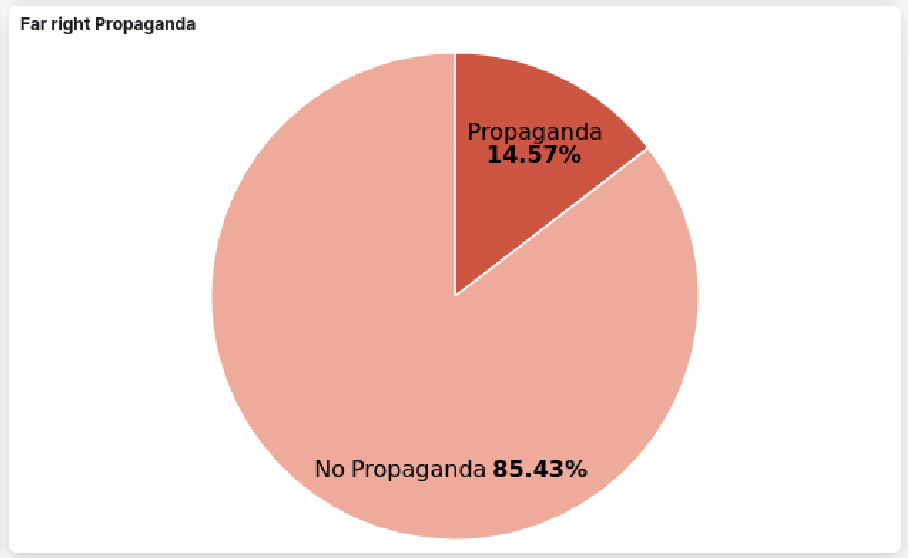


Figure 11: Percentage of propaganda in the far-right ideology for the Twitter use case

As described, using a machine learning model to predict propaganda in a text has been added to the intelligent engine. In this use case, the appearance of propaganda is studied and compared across all considered ideologies and narratives. To illustrate the usefulness of the use case and to allow its study, a dashboard Figure 11 shows a graph that is contained in the associated dashboard. More concretely, the said graph shows the percentage of messages that have been classified as propaganda. As described in D4.2, any filtering done in the dashboard would refresh this visualization.

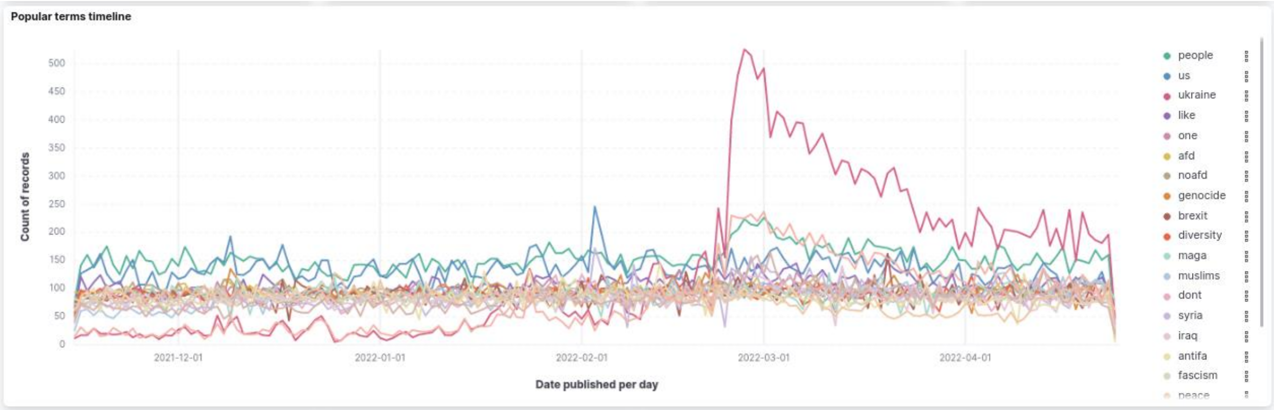


Figure 12: Visualization of the timeline of popular terms in the Twitter use case

Another interesting observation is that the dashboard allows users to identify temporal trends. For example, attending to the popular terms timeline in this use case (Figure 12), it can be seen that the term *ucraina* rises in use on the Feb. 23, 2022. This, of course, shows a general trend in the monitored users that can be explained by current geopolitical events. The hashtags used for collecting data in this use case are shown. The terms in parenthesis (e.g., (#syria, #holocaust) specify that these hashtags are captured concurrently. That is, the engine captures the messages where the hashtags co-occur.

❖ *Religious ideology*

– Pro narrative. #iraq, #islamicstate, #alleyesonisis, #syria, #khilafarestored, #islam, #muslims, #brotherhood, #MuslimLivesMatter, #UyghurGenocide, #Islamophobia, #childrenofsyria, #Uyghurs, #Gaza, #Lebanon, #Yemen, #LiberalMuslim, #MuslimLiberal, #genocide, (#muslimliberal, #feminist, #hijab), (#syria, #holocaust), (childrenofsyria, #childrenofummah, #displacedcamps), (#muslimlivesmatter, #hijabisourright)

– Counter narrative. #eurotopia, #antiterrorism, #antiterror, #peace, #antiterrorist, #againstterrorism, #stopterrorim

– Alternative narrative. #notanotherbrother, #wearethemany, #notinmyname, #(notinmyname, islam), #youAintMuslimBruv

❖ *Separatism ideology*

– Pro narrative. #indyref, #Brexit, #VoteLeave, #SpainIsAFascistState, #ANC, #separatism, #PKK

– Counter narrative. #DogsAgainstBrexit, #nationalidentity, #Framing, #Remain

– Alternative narrative. #StrongerIn

❖ Far-right ideology

– Pro narrative. #supremacy, #invasion, #GreatReplacement, #defendEurope, #Qanon, #Pizzagate, #nazism, #incel, #fascism, #antifeminist, #Gamergate, #AfD, #MAGA, #CoronaVirusDE, #SaveTheChildren, #rapefugees, #BastaLockdown, #1488, #adrenochrome, #GreatReset, #Saveourchildren, #WWG1WGA, #TaketheOath, #eurabia, #stopinvasion, #120db, #climatelockdown, #healthdictatorship, #WayfairGate, #mattonisti-user, #(rapefugees, eurabia, stopinvasion), #(MeToo,120db), #1488

– Counter narrative. #diversity, #stopHate, #antifascist, #nonazis, #FCKNZS, #noafd

– Alternative narrative. #hopenothate, #LeaveNoOneBehind

- Far-left ideology

  – Pro narrative. #socialism, #cityworkers, #1Mai, #antifa, #PariserKommune, #commune, #Commune71, #ViveLaCommune, #Ideology, #8Mai, #Marx, #Revolution, #Courage

  – Counter narrative. #AntiAntifa, #antisocialism, #GoodNightLeftSide

  – Alternative narrative. #WhitePrivilege

5.2. Monitoring Propaganda on Reddit

As described, Reddit is an interesting social network to perform studies of language use in different online communities. A use case has been defined to monitor language use in several communities that address the issue of religion. We include in this use case the analysis of potentially interesting communities such as r/HalalJihadis, r/SalafisUnveiled, and r/islamogram. Similarly, this use case studies the characteristics of the language used in these communities. For example, Figure 13 shows the relevant terms of this use case. As expected, the reader can see that most of these terms refer to religious concepts. Also, Figure 14 shows the distribution of the personal concerns detected in this use case. Again, it can be seen that religion is a predominant personal concern. Following, the subreddits used for capturing the Reddit data are shown, distributed as the engine categorizes them into the religious ideology and the different narratives.

- *Religious ideology*

  – Pro narrative. r/HalalJihadis, r/SalafisUnveiled, r/islamogram.

  – Counter narrative. r/religiousfruitcake.

  – Alternative narrative. r/islam, r/progressive Islam, r/Izlam, r/exmuslims, r/Muslim Memes, r/syriancivilwar, r/IslamicHistoryMeme, r/Hijabis, r/converts, r/MuslimLounge

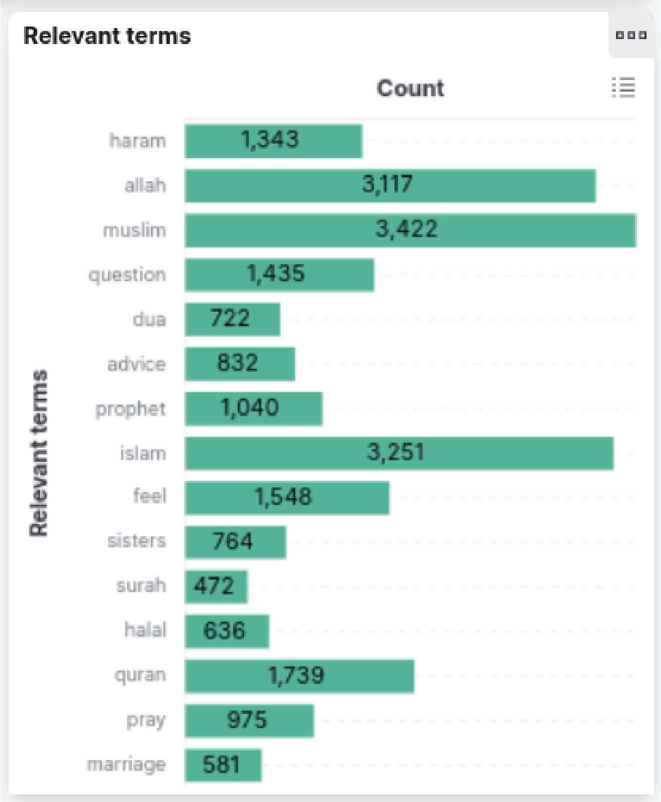


Figure 13: Visualization of the relevant terms for the Reddit use case

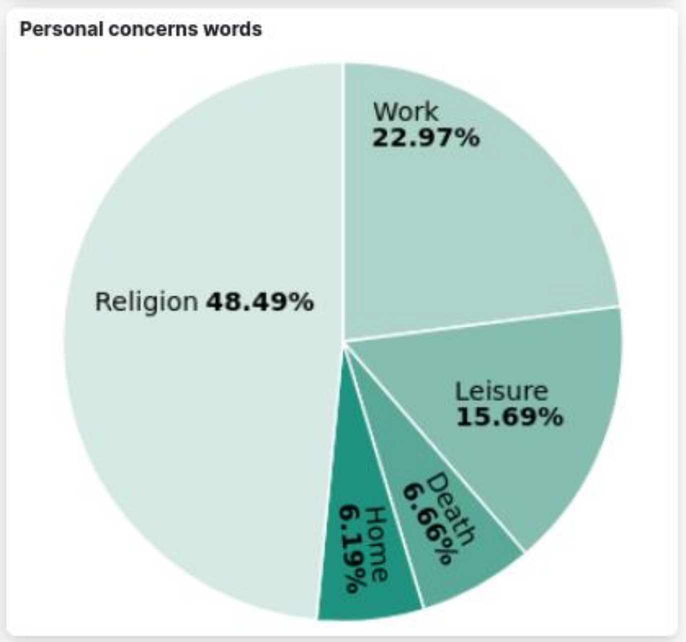


Figure 14: Visualization of the personal concerns detected in the Reddit use case

5.3 Monitoring Italian far-right

As part of the utility of the intelligent engine, it is possible to develop specific use cases that study a particularly attractive sector of users. In this way, the use case of the study of the Italian far-right language has been developed.

This use case addresses the analysis of an active group in social media, the so-called *mattonisti* (Mossetti (2021)). Their more prominent characteristic is showing a brick emoji in their username. It is a movement that belongs to Italy, thus, the Italian language has been considered in this study. Also, this is a recent movement. Therefore, it is still not clear as their objective and discourse are. Such a situation is ideal for its study with the developed intelligent engine, as its analytics offer a view of the language and social media use.

To scrape the messages from these users, we have followed a methodology in two phases. The first phase consists in obtaining users that show the brick emoji in their username. With that objective, we use the search function in the Twitter API Twitter (2021) to obtain tweets which contain said emoji characters in a given range of time of three months. As a matter of fact, the API returns messages where the emoji can appear in the text itself or the username. When obtaining a large enough set of these results (around 500,000 results), we then filter this set by removing the messages that do not have the brick emoji in the username. In this way, the system finalizes this phase with the obtained list of unique users. Following, in the second phase, we obtain the messages written by each user that appears in the obtained list of users. With this data, we can then use the data analysis and visualizations of the intelligent engine. At the time of writing, we have identified 1,631 users to monitor. These users were captured in a period between January 2022 and April 2022.

Figure 15 shows the activity of the captured *mattonisti* users. It is of special interest the time range of the activity of the captured users. As described, the user list has been compiled by capturing a large number of messages in a time range of three months. Interestingly, the time range of extraction and the activity of these users is highly overlapped. That is, the users that have generated activity in the captured time range have not been active previously. This suggests that the *mattonisti* captured accounts are of new creation, which may indicate the objective of these users. In line with this, an interesting observation that can be made is to check the proportion of bot accounts against real users.

Also, Figure 16 shows a word cloud containing the most used hashtags from the captured messages. This visualization helps to identify key topics that are discussed by this community, including current geopolitical issues such as *Ucraina*, *Russia*, and the European *greenpass*.

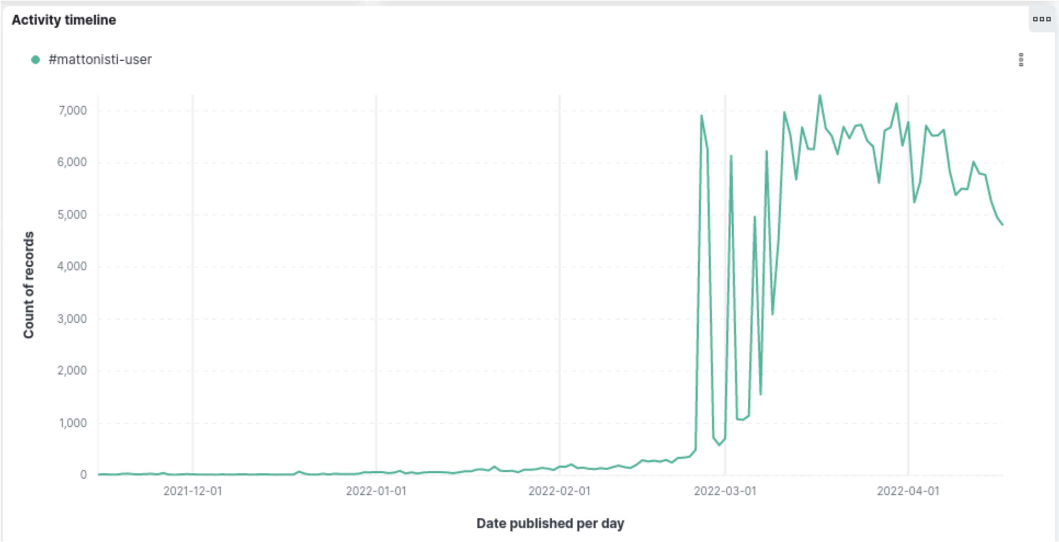


Figure 15: Activity of the captured mattonisti users

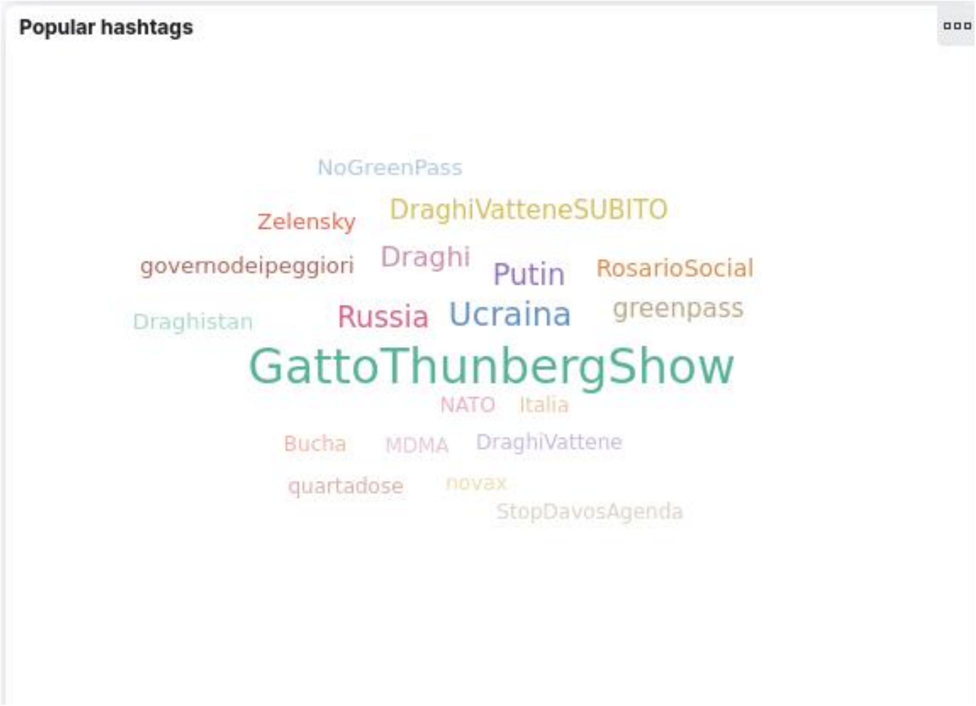


Figure 16: Wordcloud with the most used hashtags by the mattonisti community

5. Conclusions

This deliverable has presented a system aimed at classifying propaganda. Classifying propaganda is a challenging task since it depends on the available training data, which is still scarce. To overcome this issue, this work has selected several datasets and proposed a modular approach so that this work can be easily extended in the future. In addition, we have proposed a novel machine learning method that outperforms state-of-the-art results by using emotion and moral signals. The deliverable also presents the software components developed to be used in the PARTICIPATION project. The extremist monitoring dashboard has been extended to address propaganda classification, and several use cases have been developed. In addition, a Chrome plugin has been developed so that users can benefit from the potential of the developed analysis services.

References

Alam, F., Cresci, S., Chakraborty, T., Silvestri, F., Dimitrov, D., Martino, G. D. S., . . . Nakov, P. (2021). A survey on multimodal disinformation detection. arXiv preprint arXiv:2103.12541.

Apache. (2022). Apache airflow documentation. Retrieved from https://airflow.apache.org/docs/apache-airflow/stable/ (Accessed: 23 April 2022)

Araque, O., Gatti, L., & Kalimeri, K. (2020). moralstrength: exploiting a moral lexicon and embedding similarity for moral foundations prediction. *knowledge-based systems*, *191*, 105184. Retrieved from https://www.sciencedirect.com/science/article/pii/s095070511930526x doi: https://doi .org/10.1016/j.knosys.2019.105184

Araque, O., & Iglesias, C. A. (2020). deep-learning approach for classifying propaganda based on emotion signals, moral signals, and embedding similarity. *In Press*

Araque, O., Zhu, G., & Iglesias, C. A. (2019). A semantic similarity-based perspective of affect lexicons for sentiment analysis. Knowledge-Based Systems, 165 , 346-359. Retrieved from https://www.sciencedirect.com/science/article/pii/S0950705118305926 doi: https://doi.org/10.1016/ j.knosys.2018.12.005

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. Computational linguistics, 34(4), 555-596.

Balalau, O., & Horincar, R. (2021). From the stage to the audience: Propaganda on reddit. In *16th conference of the european chapter of the association for computational linguistics (eacl 2021)*.

Barrett, R. (2012). *The use of the internet for terrorist purposes*. Vienna: United Nations Office on Drugs and Crime.

Barrón-Cedeno, A., Jaradat, I., Da San Martino, G., & Nakov, P. (2019). Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management, 56* (5), 1849–1864.

Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. *In Proceedings of the 5th italian conference on computational linguistics* (pp. 1–6).

Benigni, M. C., Joseph, K., & Carley, K. M. (2017). Online extremism and the communities that sustain it: Detecting the isis supporting community on twitter. *PloS one, 12* (12), e0181405.

Chang, R.-C., & Lin, C.-H. (2021). Detecting propaganda on the sentence level during the covid-19 pandemic. arXiv preprint arXiv:2108.12269 . Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., & Guerini, M. (2019, July). CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2819– 2829). Florence, Italy: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P19-1271 (Available at https://github.com/marcoguerini/CONAN) doi: 10.18653/v1/P19-1271

Da San Martino, G., Cresci, S., Barrón-Cedeño, A., Yu, S., Di Pietro, R., & Nakov, P. (2021). A survey on computational propaganda detection. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence* (pp. 4826–4832).

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th international aaai conference on web and social media* (p. 512-515).

Decety, J., Pape, R., & Workman, C. I. (2018). A multilevel social neuroscience perspective on radicalization and terrorism. *Social Neuroscience, 13* (5), 511529. doi: 10.1080/17470919.2017.1400462

Dimitrov, D., Ali, B. B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., . . . Martino, G. D. S. (2021). Semeval-2021 task 6: Detection of persuasion techniques in texts and images. *arXiv preprint arXiv:2105.09284*. (Dataset available at https://github.com/di-dimitrov/SEMEVAL-2021-task6-corpus)

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., . . . Danforth, C. M. (2015). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences, 112* (8), 2389-2394. doi: 10.1073/pnas.1411678112

D'Ulizia, A., Caschera, M. C., ferri, F., & Grifoni, P. (2021, Mar). *Repository of fake news detection datasets*. 4TU.ResearchData. Retrieved from https://data.4tu.nl/articles/dataset/Repository_of_fake_news detection_datasets/14151755/1 doi: 10.4121/14151755.v1

Gaikwad, M., Ahirrao, S., Phansalkar, S., & Kotecha, K. (2021a). Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. *IEEE Access, 9*, 48364–48404.

Gaikwad, M., Ahirrao, S., Phansalkar, S., & Kotecha, K. (2021b). Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. *IEEE Access, 9*, 48364-48404. doi: 10.1109/ACCESS.2021.3068313

Gallagher, A., O'Connor, C., Vaux, P., Thomas, E., & Davey, J. (2021). The extreme right on discord. *Institute for Strategic Dialogue.*

Gallagher, R. J., Frank, M. R., Mitchell, L., Schwartz, A. J., Reagan, A. J., Danforth, C. M., & Dodds, P. S. (2021). Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Science*, *10* (1), 4. doi: 10.1140/epjds/s13688-021-00260-3

Grover, T., & Mark, G. (2019, Jul.). Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. *Proceedings of the International AAAI Conference on Web and Social Media, 13* (01), 193-204. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/3221

Gruppi, M., Horne, B. D., & Adalı, S. (2021). NELA-GT-2020: A large multilabelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2102.04567*.

Johnston, A., & Marku, A. (2020). Identifying extremism in text using deep learning. In *Development and analysis of deep learning architectures* (pp. 267–289). Springer.

Johnston, A. H., & Weiss, G. M. (n.d.). Identifying sunni extremist propaganda with deep learning. *In 2017 ieee symposium series on computational intelligence (ssci).*

Kausar, S., Tahir, B., & Mehmood, M. A. (2020). Prosoul: a framework to identify propaganda from online urdu content. *IEEE Access, 8*, 186039–186054.

Kayode-Adedeji, T., Oyero, O., & Aririguzoh, S. (2019). Dataset on online mass media engagements on youtube for terrorism related discussions. *Data in brief, 23*, 103581. (Available at https://www.sciencedirect.com/science/article/pii/S2352340918315592#ec0005)

Martino, G. D. S., Barrón-Cedeño, A., & Nakov, P. (2019). *Findings of the nlp4if-2019 shared task on fine-grained propaganda detection*. (Dataset available at https://propaganda.qcri.org/nlp4if-shared-task/data/)

Martino, G. D. S., Cresci, S., Barrón-Cedeño, A., Yu, S., Pietro, R. D., & Nakov, P. (2020). A survey on computational propaganda detection. *CoRR, abs/2007.08024*. Retrieved from https://arxiv.org/abs/2007.08024

Moral, P., Marco, G., Gonzalo, J., Carrillo-de-Albornoz, J., & Gonzalo-Verdugo, I. (2023). Overview of DIPROMATS 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers. Procesamiento del lenguaje natural, 71, 397-407.

Mossetti, P. (2021). Come l'emoji del mattone è diventata un simbolo di destra su twitter. *Wired*. Retrieved from https://www.wired.it/internet/social -network/2021/03/03/emoji-mattone-simbolo-destra-twitter/

Moves, P. (2022). Extremesentilex. Retrieved from http://moves.di.ubi.pt/extremesentilex.html (Accessed: 23 April 2022)

O'Connor, C. (2021). The extreme right on twitch. *Institue for Strategic Dialogue*.

Pedersen, T. (2001). A decision tree of bigrams is an accurate predictor of word sense. *arXiv preprint cs/0103026*.

Pramanick, S., Dimitrov, D., Mukherjee, R., Sharma, S., Akhtar, M., Nakov, P., . . . others (2021). Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413*. (Dataset available at https://github.com/di-dimitrov/harmeme)

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political factchecking. *In Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2931–2937). (Dataset available at https://hrashkin.github.io/factcheck.html)

Richardson, A. J. (2020). *Estimating the impact of political propaganda on reddit users' political opinions* (Unpublished doctoral dissertation). Georgetown University.

Saini, J. K., & Bansal, D. (2021). Detecting online recruitment of terrorists: towards smarter solutions to counter terrorism. *International Journal of Information Technology, 13* (2), 697–702.

Sánchez-Rada, J. F., & Iglesias, C. A. (2019, December). Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison. Information Fusion, 52, 344-356. Retrieved from https:// www.sciencedirect.com/science/article/pii/S1566253518308704

Scanlon, J. R., & Gerber, M. S. (2014). Automatic detection of cyberrecruitment by violent extremists. *Security Informatics, 3* (1), 1–10.

Seo, H. (2014). Visual propaganda in the age of social media: An empirical analysis of twitter images during the 2012 israeli–hamas conflict. *Visual Communication Quarterly, 21* (3), 150–161.

Tundis, A., Mukherjee, G., & Mühlhäuser, M. (2020). Mixed-code text analysis for the detection of online hidden propaganda. In *Proceedings of the 15th international conference on availability, reliability and security* (pp. 1–7).

Twitter. (2021). Twitter developer platform. Retrieved from https://developer.twitter.com/en/docs (Accessed: 13 July 2021)

van der Vegt, I., Mozes, M., Kleinberg, B., & Gill, P. (2021). The grievance dictionary: understanding threatening language use. *Behavior research methods*, 1–15

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Volkova, S., Ayton, E., Arendt, D. L., Huang, Z., & Hutchinson, B. (2019). Explaining multimodal deceptive news prediction models. In *Proceedings of the international aaai conference on web and social media* (Vol. 13, pp. 659– 662).

Von Eye, A., & Mun, E. Y. (2014). Analyzing rater agreement: Manifest variable methods. Psychology Press.

Wang, S., & Manning, C. (2012, July). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th annual meeting of the association for computational linguistics (volume 2: Short papers*) (pp. 90– 94). Jeju Island, Korea: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P12-2018

Participation