Information Processing and Management 56 (2019) 1849–1864

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Proppy: Organizing the news based on their propagandistic content

Alberto Barrón-Cedeño[*,1,a], Israa Jaradat[1,b], Giovanni Da San Martino[c], Preslav Nakov[c]

[a] Università di Bologna, Forlì, Italy
[b] University of Texas at Arlington, USA
[c] Qatar Computing Research Institute, HBKU, Qatar

ARTICLE INFO

2010 MSC:
00-01
99-00

Keywords:
Propaganda detection
News bias
Investigative journalism

ABSTRACT

Propaganda is a mechanism to influence public opinion, which is inherently present in extremely biased and fake news. Here, we propose a model to automatically assess the level of propagandistic content in an article based on different representations, from writing style and readability level to the presence of certain keywords. We experiment thoroughly with different variations of such a model on a new publicly available corpus, and we show that character *n*-grams and other style features outperform existing alternatives to identify propaganda based on word *n*-grams. Unlike previous work, we make sure that the test data comes from news sources that were unseen on training, thus penalizing learning algorithms that model the news sources used at training time as opposed to solving the actual task. We integrate our supervised model in a public website, which organizes recent articles covering the same event on the basis of their propagandistic contents. This allows users to quickly explore different perspectives of the same story, and it also enables investigative journalists to dig further into how different media use stories and propaganda to pursue their agenda.

1. Introduction

The landscape of news outlets is wide: from supposedly neutral to clearly biased. When reading a news article, every reader should be aware that, at least to some extent, it inevitably reflects the bias of both the author and the news outlet where the article is published. However, it is difficult to identify exactly what the bias is. It could be that the author herself may not be conscious about her own bias. Or it could be that the article is part of the author's agenda to persuade readers about something on a specific topic. The latter situation represents propaganda. According to the now classical work from the (Institute for Propaganda Analysis, 1938), propaganda can be defined as follows:

Definition 1. Propaganda is expression of opinion or action by individuals or groups deliberately designed to influence opinions or actions of other individuals or groups with reference to predetermined ends.

Propaganda is most effective when it can go unnoticed. That is, if a person reads a *journalistic* text, in a formal or an informal news outlet (e.g., in a blog or in social media) she should not be able to identify it as propagandistic. In that case, the reader is

* Corresponding author.
E-mail addresses: a.barron@unibo.it, albarron@gmail.com (A. Barrón-Cedeño), israa.jaradat@mavs.uta.edu (I. Jaradat), gmartino@hbku.edu.qa (G. Da San Martino), pnakov@hbku.edu.qa (P. Nakov).
[1] Work carried out mostly while at the Qatar Computing Research Institute, HBKU.

https://doi.org/10.1016/j.ipm.2019.03.005
Received 25 July 2018; Received in revised form 24 January 2019; Accepted 12 March 2019
Available online 16 May 2019
0306-4573/

exposed to the propagandistic content without her knowledge and some of her opinions might change as a result. A striking example of the use of propaganda was allegedly put in place to influence the 2016 US Presidential elections (Muller, 2018). Given the wide landscape of news outlets —from tabloids to broadsheets, from printed to digital, from objective to biased— we believe that both news consumers and institutions might benefit from an automatic tool that can detect propagandistic articles.

Here we propose `proppy`, a system to organize news events according to the level of propagandistic contents in the articles covering them. `Proppy` is a full architecture (cf. Fig. 3) that takes a batch of news articles as input, identifies the covered events, and organizes each event according to the level of propaganda in each article. Our major contribution, and the focus of this manuscript, is a supervised model to compute what we refer to as *propaganda score*: the estimated likelihood of a text document to contain propagandistic mechanisms to deliberately influence the reader's opinion.

`Proppy` computes a propaganda score using a maximum entropy classifier. We chose this classifier in order to facilitate direct comparison to previous work (Rashkin, Choi, Jang, Volkova, & Choi, 2017) and to focus our efforts on improving the representation of the data in terms of features. In Rashkin et al. (2017), word *n*-grams were used but, as the authors themselves pointed out, this yielded significant drop in performance when testing on articles from sources that were not seen on training. Here we aim to shed some light about why this could be the case. Therefore, we formulate the following hypothesis:

Hypothesis 1 (*H1*). Representations based on writing style and readability can generalize better than currently-used approaches based on word-level representations.

We argue that this is because word-level representations tend to learn topic and source, rather than whether the target article is propagandistic or not. In order to test the above hypothesis, we first replicated a pre-existing model for propaganda detection (Rashkin et al., 2017).[2] Later on, we compiled a new corpus —`QProp`— which, unlike most pre-existing corpora, keeps explicit information about the source of each article, thus allowing us to train on articles from some sources and to test on articles from different sources that have not been used for training. We design experiments that involve training and evaluating several supervised models using features based on text readability and style; such features have been widely used in authorship attribution tasks (Stamatatos, 2009). In our thorough experimentation, we obtain statistically significant improvements over existing approaches in terms of classification performance, especially when testing on articles from unseen sources.

Our contributions can be summarized as follows:

1. We experiment with different families of feature representations (some of them used for the first time for this task) spanning readability, vocabulary richness, and style in an effective propaganda estimation model, and we demonstrate empirically that they are effective for actually detecting propaganda, as opposed to learning the article's source or its topic as it is the case in most previous work.
2. We release a new dataset of 51k full-text news articles,[3] together with the source code of our implementation. Unlike previous datasets, for each article, we provide metadata including the source and whether it is considered propagandistic.
3. We release a webapp that allows users to explore the coverage of the current news events on the basis of their propagandistic content.[4]

The remainder of this article is organized as follows. Section 2 offers a soft introduction to propaganda. Section 3 presents related work on (automatic) propaganda identification and authorship-derived representations. Section 4 introduces our propaganda detection model. Section 5 presents the datasets we experiment with, including our new dataset. Section 6 covers our experiments and discusses the results. Section 7 describes the full architecture of `proppy` —as running on the Web— which includes retrieving the articles, grouping them into events, computing their propaganda score, and displaying the results. Finally, Section 8 concludes and points to possible directions for future work.

2. Background

The term *propaganda* was coined in the 17th century, meaning propagation of the Catholic faith (Jowett and O'Donnell, 2012, p. 2). The term soon took a pejorative connotation, as it was not only intended to spread the faith in the New World, but also to oppose Protestantism; i.e. it was not neutral. Here, we are interested in a journalistic point of view of propaganda: how news management lacking neutrality shapes information by emphasizing positive or negative aspects purposefully (Jowett and O'Donnell, 2012, p. 1). As Jowett and O'Donnell mention, propaganda is frequently considered a synonym of lies, distortion, and deceit (Jowett and O'Donnell, 2012, p. 2). Indeed, all biased messages have been identified as propagandistic, regardless of whether the bias was conscious or not (Ellul, 1965, p. XV). As a result, if a model is capable of identifying propaganda in a piece of news, it enhances a reader's awareness that she might be facing a biased text. Bias must be considered when addressing people's information needs, as it affects us all and much of the time we are unaware of it (Baeza-Yates, 2018).

One of the seminal categorizations of propaganda devices dates back to the 1930s. The 1936 US election campaign between then

[2] That model further predicts different kinds of bias and factuality of reporting, which are beyond the scope of this article.
[3] The code and the dataset are available at http://proppy.qcri.org/about.html
[4] The website is accessible at http://proppy.qcri.org

President Franklin D. Roosevelt and Alf Landon attracted the attention of scholars to the language used by the contenders.[5] Clyde R. Miller proposed one of the seminal categorizations of propaganda in 1937.[6] It consists of seven devices (Institute for Propaganda Analysis, 1938), which remain well accepted today (Jowett and O'Donnell, 2012, p.237)[7]:

1. **Name calling** appeals to hate and fear by giving "bad names" to individuals, groups, nations, races, policies, beliefs, and ideals to make the reader condemn or reject them.
2. **Glittering generalities** identify a message with virtue by using "virtue words" that appeal to emotions of love, generosity, and brotherhood (e.g., freedom, honor, liberty, progress).
3. **Transfer** carries the authority, sanction, and prestige of something we respect/revere to have us accept something we would not otherwise.
4. **Testimonial** use authoritative persons' testimonials (e.g., celebrities, experts, public figures), such as quotes, to strengthen an argument.
5. **Plain folks** try to win confidence by appearing to be common people as ourselves (e.g., a politician using simple and friendly language during a campaign).
6. **Card stacking** involves stacking cards against the truth by under/over-emphasizing to dodge the issues and to evade the facts (it implies lying, omitting facts, and offering false testimonies).
7. **Band wagon** appeals to groups held together by a common tie (e.g., nationality, religion, gender) in order to push them to follow the crowd —as *the crowd is always right*.

Regardless of the particularities of each device, they all have a common objective: to make a person adopt a judgment or feeling without examining the actual evidence. Such devices are used to further political ambitions, to attract support for questionable policies, and to deflect the attention from the real issues (Bazerman, 2010, p. 106). Be it fake news, bias, or other resembling phenomena, the devices of propaganda are commonly used to transmit the message through emotions that blurry the judgment of the receptor (Institute for Propaganda Analysis, 1938).

Some devices can be easily spotted in written media —with an educated eye. As the examples in Fig. 1 show, this is the case for devices such as *name calling* and *glittering generalities*. Clearly positive or negative words and phrases, together with some keywords, trigger the alarms. Devices such as *card stacking* require to grab evidence from external sources to find out whether information is being hidden or it is instead irrelevant. Here, we focus on propaganda that can be spotted intrinsically. That is, by analyzing a document in isolation.

3. Related work

Recently, there has been a lot of interest in studying disinformation and bias in the news and in social media. This includes challenging the truthiness of news (Brill, 2001; Finberg, Stone, & Lynch, 2002; Hardalov, Koychev, & Nakov, 2016; Potthast, Kiesel, Reinartz, Bevendorff, & Stein, 2018), of news sources (Baly, Karadzhov, Alexandrov, Glass, & Nakov, 2018), and of social media posts (Canini, Suh, & Pirolli, 2011; Castillo, Mendoza, & Poblete, 2011; Zubiaga, Liakata, Procter, Wong Sak Hoi, & Tolmie, 2016), as well as studying credibility, influence, and bias (Ba, Berti-Equille, Shah, & Hammady, 2016; Baly et al., 2018; Chen, Wu, Srinivasan, & Zhang, 2013; Kulkarni, Ye, Skiena, & Wang, 2018; Mihaylov, Georgiev, & Nakov, 2015; Mihaylov et al., 2018). The interested reader can also check several recent surveys that offer a general overview on "fake news" (Lazer et al., 2018), or focus on topics such as the process of proliferation of true and false news online (Vosoughi, Roy, & Aral, 2018), on fact-checking (Thorne & Vlachos, 2018), on data mining (Shu, Sliva, Wang, Tang, & Liu, 2017), or on truth discovery in general (Li et al., 2016). For some specific topics, research was facilitated by specialized shared tasks such as the SemEval-2017 task 8 on Rumor Detection (Derczynski et al., 2017), the CLEF-2018 lab on Automatic Identification and Verification of Claims in Political Debates (Nakov et al., 2018), the FEVER-2018 task on Fact Extraction and VERification (Thorne, Vlachos, Christodoulopoulos, & Mittal, 2018), and the SemEval-2019 Task 8 on Fact Checking in Community Question Answering Forums (Mihaylova et al., 2019), among others.

From a modeling perspective, most approaches relied on stylistic and complexity representations, which tend to be topic- and genre-independent. That is, regardless of the event being covered in the target news article or the direction of its bias (if any), the features need to contain the necessary information for the model to be able to make a decision. This is precisely the main design principle of the representations used in authorship attribution —the task of verifying whether a dubious text has been written by the same known author who is behind a number of other texts (Juola, 2012). While factors such as topic and text length play little role for this task, among the most successful representations we typically find character-level *n*-grams (Stamatatos, 2009). As Hypothesis 1 states, we believe that these representations are robust and are also useful for modeling the degree of bias and propaganda in news articles.

[5] This is to the *fake news* and *the post-truth age* phenomena, which spawned during the 2016 US presidential campaign (Davies, 2016).

[6] The devices were published in an unsigned article. Recent studies identify Miller as the creator (Sproule, 2001).

[7] For simplicity, we opt to stick to this categorization even if some recent ones include many more devices. Other scholars consider categorizations with as many as eighty-nine techniques (Conserva, 2003), just to give an example. At the time of writing (July 2018), the Wikipedia article on propaganda techniques had 60+ categories (http://en.wikipedia.org/wiki/Propaganda_techniques). In general, the devices in these categorizations are subtypes of the general schema proposed in Institute for Propaganda Analysis (1938).

Manchin says Democrats acted like **babies**$_1$ at the SOTU

In a glaring sign of just how **stupid**$_1$ and **petty**$_1$ things have become in Washington these days, Manchin was invited on Fox News Tuesday morning to discuss how he was one of the only Democrats in the chamber for the State of the Union speech not looking as though Trump **killed his grandma**$_1$. When others in his party declined to applaud even for the most uncontroversial of the president's remarks, Manchin did. **He even stood for the president when Trump entered the room, a customary show of respect for the office in which his colleagues declined to participate**$_2$.

"**That's the way I was raised in West Virginia. We have respect**$_3$," he said when asked why he didn't follow Nancy Pelosi's lead. [...]

The Democrats show on Tuesday illustrates just how far the party is willing to go to avoid working with Trump, even as his policies continue to **improve the economy**$_4$ and set the stage for **major infrastructure improvements**$_4$ in the years ahead. [...]

1. **Name calling**. Loaded words are intended to exaggerate the claimed lack of respect in the act.
2. **Card stacking**. The author is overemphasizing by hiding information: graphic evidence shows that Mr. Manchin was not the only Democrat standing.
3. **Glittering generalities**. Everybody in a region shares the "same right attitude".
4. **Glittering generalities**. The author gives virtue to the presidential projections for progress.

Fig. 1. Extracts of a news article with propaganda devices identified (explanation at the bottom). Article published by Personal Liberty on January 31st 2018 (last visit: April 15th, 2019). http://personalliberty.com/manchin-says-democrats-acted-like-babies-sotu/.

Previous work on bias and disinformation detection has already looked into such kinds of features. In their efforts to assess the credibility of claims, (Popat, Mukherjee, Strötgen, & Weikum, 2017) considered what they call *stylistic features* —occurrence of assertive and factive verbs, hedges, implicative words, report verbs, and discourse markers—, which they extracted using manually crafted gazetteers. In contrast, our main focus is on style markers such as sequences of characters and readability measures.

Stylometry has been considered when looking for hyper-partisanship, i.e., *extremely one-sided news* such as extreme-left and extreme-right. For example, Potthast, Kiesel, Reinartz, Bevendorff, and Stein (2017) used articles from nine sources, whose factuality had been manually verified by professional journalists. Their interest was in identifying true news vs. satire vs. "fake news." For that, they applied a stylometric analysis, which was originally designed for authorship verification (Koppel, Schler, & Bonchek-Dokow, 2007), to predict factuality (fake vs. real) and bias (left vs. right vs. mainstream; or hyperpartisan vs. mainstream). Their hypothesis —similar to ours—, was that biased texts share writing style, regardless of their political preferences or, in general, topic and conveyed message. In order to characterize the writing style, (Potthast et al., 2017) considered character *n*-grams, stopwords, and part-of-speech tags together with a number of readability scores. They also looked at specific words and the average length of the paragraphs in the texts, among other domain-specific features. They observed that the writing style of left- and right-biased texts is very similar; nevertheless, they found out that their representations were not effective enough for the task of "fake news" identification. One danger was that their classifier could learn the (style of the) *source* rather than the actual task; they addressed this by discarding all features occurring in less than 10% of the corpus. In contrast, here we play with the data distribution, making sure that the test data comes from news sources that were unseen on training, thus penalizing models that learn to predict the (style of) the specific news sources used at training time as opposed to solving the actual task.

Writing style and complexity were considered in the efforts carried out by Horne and Adal (2017) to differentiate real news vs. "fake news" vs. satire. Among the stylistic markers they used, they included the number of occurrences of different part-of-speech tags, swearing and slang words, stopwords, punctuation, and negation. Regarding complexity, they considered various readability measures (we use most of these measures as well; cf. Table 3). According to their analysis, "fake news" tend to be shorter and use simpler language, i.e. shorter and less technical words and their readability index is lower. Thus, they concluded that, under their representation, "fake news" are closer to satire. Even though they considered more than 130 features, they did not use them in a learning algorithm and did not perform a systematic evaluation. Here, we include their features in our experimentation and we compare them with other kinds of features.

Rashkin et al. (2017) aimed at differentiating real news from satire, hoaxes, and propaganda. In order to do that, they compiled a corpus of documents from the English Gigaword (real news) and from seven unreliable news sources for the three other categories. Their representation was based on word *n*-grams, with $n \in [1, 3]$. We use their corpus in our experiments (cf. Section 5.1) and we consider their representations as a baseline (cf. Section 6). As Rashkin et al. pointed out, the use of word *n*-grams yielded significant drop in performance when testing on news articles from sources unseen on training as opposed to testing on seen sources. Here, we evaluate the hypothesis that this drop is due to word *n*-grams being topic-dependent and modeling the news source rather than the concept of propaganda. We further propose features that can overcome these limitations.

Finally, there have been efforts by communities of experts in journalism to raise awareness by evaluating the contents published by

different news outlets and in social media. Below we mention some of these efforts. For instance, the *CrossCheck* project has set up an infrastructure to identify popular content in social media and to compare it against the coverage by traditional outlets.[8] *Full Fact*,[9] an independent fact-checking organization in the UK, provides free tools, information and advice for checking claims by politicians and the media. Other popular fact-checking organizations include Snopes,[10] Politifact,[11] and FactCheck.[12] *Media Bias/Fact Check*[13] (MBFC) gathers volunteers who, among other activities, measure the bias of entire news sources. For each source, MBFC provides a bias score, factuality of reporting, and URL, among other information. It further provides a curated list of questionable sources —some of which are flagged as propagandistic. We depart from MBFC judgments on propaganda to build our corpus, as explained in Section 5.2 below.

4. Representations

We use a maximum entropy classifier with L2 regularization and default parameters to discriminate propagandistic from non-propagandistic articles. This is the same classifier as the one used by Rashkin et al. (2017), and we chose it in order to facilitate direct comparison with their work. We consider four families of features, which we describe below.

4.1. Word n-gram features

We use $tf - idf$-weighted word [1,3]-grams as baseline features, after tokenizing the text with NLTK (Bird, Loper, & Klein, 2009). They were used in Rashkin et al. (2017) to discriminate trusted vs. propaganda vs. hoax vs. satire articles (cf. Section 3).

4.2. Lexicon features

As discussed in Section 2, certain kind of vocabulary is common for specific propagandistic techniques (e.g., in *name calling* and *glittering generalities*). We try to capture this by considering representations reflecting the frequency of specific words from a number of lexicons, shown in Table 1. They come from the Wiktionary, the Linguistic Inquiry and Word Count (LIWC) lexicon (Pennebaker, Francis, & Booth, 2001), Wilson's subjectives (Wilson, Wiebe, & Hoffmann, 2005), Hyland hedges (Hyland, 2015), and Hooper's assertives (Hooper, 1975). For each of the 18 lexicons, we count the total number of occurrences of the words from this lexicon in the text.

Rashkin et al. (2017) studied the relationship between the occurrence of the words from the above lexicons in different kinds of news articles. They found that the words from some of their lexicons (e.g., *swear, see, negation*) appear more frequently in propagandistic, satire, and hoax articles than in trustworthy news articles.

4.3. Vocabulary richness, readability, and style

Potthast et al. (2017) showed that hyperpartisan outlets tend to use writing style that is different from that of mainstream news media. Thus, we also use features that model style. Different topic-independent features have been proposed in the literature to characterize the vocabulary richness, style, and complexity of a text. Whereas many such features were originally intended to assess the pertinence of teaching materials for different education levels, they have been also found useful for authorship attribution and related tasks (Stamatatos, 2009). Table 2 shows the five features we use in order to model the vocabulary richness of a news article. We consider the *type-token ratio* (TTR) as well as the number of types appearing exactly once or exactly twice in the document: the *hapax legomena* and *dislegomena*, respectively. We further combine word types, word tokens, and hapax legomena to compute Honore's R (Honore, 1979) and Yule's characteristic K (Yule, 1944).

Table 3 shows the three readability features: the *Flesch–Kincaid grade level* (Kincaid, Fishburne, Rogers, & Chissom, 1975), the *Flesch reading ease* (Kincaid et al., 1975), and the *Gunning fog index* (Gunning, 1968).

Stamatatos (2009) argues that in tasks in which the topic is not relevant, character-level representations are more sensitive than token-level ones. He considers that "the most frequent character *n*-grams are the most important features for stylistic purposes". Our style representation consists of $tf - idf$-weighted character 3-grams. These representations capture different style markers, such as prefixes, suffixes, and punctuation marks.

4.4. Nela

Recently, Horne, Khedr, and Adal (2018) presented the NEws LAndscape features (NELA): 130 content-based features collected from the literature that measure different aspects of a news article such as sentiment, bias, morality, and complexity, among others. We integrated the NELA features into our model and experiments. They are categorized in six subgroups, which are included in

[8] http://firstdraftnews.org/crosscheck-qualitative-research.
[9] http://fullfact.org/
[10] http://www.snopes.com/
[11] http://www.politifact.com/
[12] http://www.factcheck.org/
[13] http://mediabiasfactcheck.com

Table 1
Lexicon sources and lexicons we use for feature extraction with example entries.

Source	Lexicon (example entry)
Wiktionary	**Modal** (truly) • **Action** (accidentally) • **Manner Adverbs** (foolishly) • **Comparative** (higher) • **Superlative Forms** (worst)
LIWC (Pennebaker et al., 2001)	**First Person Singular** (my) • **Second Person** (you) • **Hear** (says) • **Money** (costs) • **Negation** (can't) • **Number** (quarter) • **See** (watch) • **Sexual** (gay) • **Swear** (dumb)
Wilson et al. (2005)	**Strong subjectives** (anti-semites) • **Weak subjectives** (extremist)
Hyland (2015)	**Hedges** (perhaps)
Hooper (1975)	**Assertives** (certain)

Table 2
Vocabulary richness features.

Feature	Computation
TTR. Type–token ratio.	$\lvert types\rvert / \lvert tokens\rvert$
Hapax legomena. Word types appearing once in a text.	$\lvert types_1\rvert$
Hapax dislegomena. Word types appearing twice in a text	$\lvert types_2\rvert$
Honore's R. Word types, tokens, and hapax legomenæ.	$\frac{100 \cdot \log(\lvert tokens \rvert)}{1 - \frac{\lvert hapax\_legomena \rvert}{\lvert types\rvert}}$
Yule's characteristic K. Combination of types appearing with different frequencies and tokens. Assumes that the occurrences of a word follow a Poisson distribution. Here $i = [1, 2, ...]$ is the number of word types with a frequency of i in the text.	$10^4 \frac{\sum_i i^2 \lvert types_i \rvert - \lvert tokens\rvert}{\lvert tokens\rvert^2}$

Table 3
Readability features.

Feature	Computation
Flesch–Kincaid grade level. US grade level necessary to understand the text.	$0.39 \cdot \frac{\lvert tokens\rvert}{\lvert syllables\rvert} + 11.9 \cdot \frac{\lvert syllables\rvert}{\lvert tokens\rvert} - 15.59$
Flesch reading ease. A scale in the range [0,100] representing the complexity of a text. Higher score means easier text.	$206.835 - 1.015 \cdot \frac{\lvert tokens\rvert}{\lvert sentences\rvert} - 84.6 \cdot \frac{\lvert syllables\rvert}{\lvert tokens\rvert}$
Gunning fog index. Number of years of formal education necessary to understand the text. Here, $tokens_c$ stands for complex tokens: those with three syllables or more.	$0.4\left(\frac{\lvert tokens\rvert}{\lvert sentences\rvert} + 100 \cdot \frac{\lvert tokens_C\rvert}{\lvert tokens\rvert}\right)$

Table 4
NELA features (Horne et al., 2018).

Subgroup	Description
Structure	part-of-speech (normalized counts)
Sentiment	emotion: positive, negative, affect, etc. from LIWC • happiness score
Topic-specific	biological process • relativity: motion, time, and space words • personal concerns: work, home, leisure, etc. (all from LIWC)
Complexity	SMOG readability measure • average word length • word count • cognitive process words from LIWC
Bias	several bias lexicons • subjectivity probability in the text
Morality	features based on the Moral Foundation Theory (Graham, Haidt, & Nosek, 2009)

Table 4 (a seventh subgroup *Facebook engagement* reported in Horne et al. (2018) was not included in their software release).

5. Corpora

We use two corpora in our experiments. In Section 5.1, we introduce the corpus created by Rashkin et al. (2017), while in Section 5.2, we present `QProp`, our new corpus, which is tailored for the kind of analysis we want to do.[14]

5.1. `TSHP-17` corpus

We refer to this corpus as `TSHP-17`. This stands for *trusted, satire, hoax and propaganda 2017* corpus —the four represented classes (Rashkin et al., 2017). Articles from eleven news outlets were considered, and the classes were assigned to news articles according to the class of the news outlet they come from. The labels for the news outlets in turn come from *US News & World Report*,[15] which uses the following four labels: *trusted, satire, hoax*, and *propaganda*. Note that there were only 1–3 sources for each class, which

[14] `QProp` is available for download at http://proppy.qcri.org/about.html.
[15] www.usnews.com/news/national-news/articles/2016-11-14/avoid-these-fake-news-sites-at-all-costs.

can easily confuse a classifier to see this as a kind of source prediction task (even though the dataset does not indicate the article source). Moreover, some of the *satire* sources only contribute a small number of articles.

Note also that the source of the trusted articles is Gigaword,[16] which includes articles from four sources. While it is unclear how the sampling in TSHP-17 was carried out, we believe (Rashkin et al., 2017) includes instances from all of them.

TSHP-17 includes a total of 22,580 news articles, and it is fairly balanced between the classes. Table 5 shows the statistics about the data distribution, including the number of articles for each of the four classes in the different partitions (train, dev, test), as well as the average length of the included articles.

We performed a number of automatic checks on TSHP-17 and we found that a number of propagandistic instances contained titles of Youtube videos and lacked textual content, e.g., *"its like this kidyoutube", "7/7 ripple effect 2 - traileryoutube"*. As a result of discarding such instances, we filtered out 330 entries from the training and 80 entries from the development sets, and we ended up with 22.5k instances. Note that this number is far from the 75k articles reported in Rashkin et al. (2017, Table 1); however, the TSHP-17 corpus which was used for the experiments in Rashkin et al. (2017), contained only 22,580 articles.

5.2. QProp corpus

TSHP-17 does not provide information about the source of the individual news articles. Therefore, we do not know which propagandistic articles were published by, e.g., *The Natural News* or *DC Gazette*; this applies to all four classes. Moreover, only a small number of sources have been used: eleven overall, only two of which are propagandistic. As a result, it is impossible to perform extensive experiments and analysis that take the source into account.[17]

Therefore, we compiled a new corpus, QProp, which stands for *QCRI's propaganda* corpus. We focused on two classes only: *propaganda* vs. *trustworthy*. We compiled QProp using information about entire news outlets as published by Media Bias/Fact Check (MBFC; cf. Section 2). We used the propaganda and the trustworthiness judgments by MBFC as labels for all articles in the respective news outlet. We considered 104 sources to download news articles from. For the *propaganda* class, we considered ten different sources. For the *trustworthy* class, we used 94 sources with different MBFC-derived bias levels: left biased, left-center biased, least biased, right-center biased, and right biased. Table 6 lists the propagandistic sources we used.

Given a target news outlet, we crawled its Web site to retrieve actual news articles, assigning to all of them the *propaganda* label from MBFC for that Web site. For the purpose, we used GDELT,[18] a real-time database with information from news outlets from all over the world (Leetaru & Schrodt, 2013). We considered the period from October, 2017 till December, 2018. In addition to the article's text and title, we further include metadata and class labels. GDELT offers rich metadata for each article, and we retrieve and include in the corpus the following information it offers: geographical information, average sentiment, publication date, identifier, author, and official source name. We further added two labels from MBFC: bias (e.g., left, right) and propaganda label (true or not); both derived from the publisher profile.

Table 7 shows statistics about QProp, including distribution for the individual classes. The corpus consists of 51.3k articles: 5.7k from propagandistic sources and 45.6k from trustworthy ones. We randomly split these articles into train/dev/test partitions with the constraint of preserving the class and the source distributions: namely, we distribute the articles from each news source into the training, development, and test partitions in a proportion of 70%, 10%, and 20%, respectively.

QProp offers a number of advantages over previous corpora. First, it includes the source of each article. This is essential as it enables us to test our hypothesis that supervised models trained to detect propagandistic content could instead be learning the news source. Second, it is more realistic, including several news sources for each of the classes.[19]

6. Experiments and evaluation

We designed three experiments to verify hypothesis H1. The first one aims at comparing our features with the ones used in Rashkin et al. (2017), and thus we experimented with a 4-way classifier: *trusted* vs. *propaganda* vs. *hoax* vs. *satire*. The second experiment focuses on our main 2-way classification task: *propaganda* vs. *non-propaganda*. We perform this experiment on both the TSHP-17 and the QProp corpora. As we observe a sizable drop in performance when testing on news coming from sources never seen during training, we further run a third experiment to test whether this is due to representations misleading the algorithm to model the media source instead of solving the actual task.

We replicate the experimental setup of Rashkin et al. (2017) by using a Maximum Entropy classifier with L2 regularization and default parameters (C = 1). This allows us to compare to them directly, and to focus on the effectiveness of the different representations: word *n*-grams, lexicon, vocabulary richness, readability, and character *n*-grams (cf. Section 4). Note that, since we fixed the hyper-parameters of the algorithm, there is no need for a separate tuning dataset. We also tried using support vector machines (Cristianini & Shawe-Taylor, 2000). The results with the linear kernel varied slightly with respect to the Maximum Entropy classifier and they were much worse when using the polynomial and RBF kernels. Thus, we decided to report results for the Maximum

[16] http://catalog.ldc.upenn.edu/LDC2003T05

[17] Besides contacting the authors, we tried to find the source of each article using different online search engines. Most articles in the TSHP-17 corpus —particularly those from the *propaganda* and the *hoax* classes—, could not be found and seem to have been removed from the Web.

[18] GDELT Project: http://www.gdeltproject.org/

[19] The source code for generating a new version of the corpus is available at http://proppy.qcri.org/about.html

Table 5
Statistics about the TSHP-17 corpus (Rashkin et al., 2017), including the number of articles and the average length (word tokens) for each of the four classes in the different partitions (train, dev, test).

Source	# Sources	# Articles	Train	Dev	Test	Length (tokens)
Trusted	4*	5,750	3,997	1,003	750	522 ± 429.13
Satire	3	5,750	3,981	1,019	750	324 ± 276.31
Hoax	2	5,750	4,014	986	750	262 ± 300.92
Propaganda	2	5,330	3,670	910	750	1,047 ± 1,156.87
Total	**11**	**22,580**	**15,662**	**3,918**	**3,000**	**529 ± 705.34**
Sources	Trusted	Gigaword News*				
	Satire	The Onion • The Borowitz Report • Clickhole				
	Hoax	American News • DC Gazette				
	Propaganda	The Natural News • Activist Report				

Table 6
Overview of the propagandistic news outlets and the number of articles they contribute to QProp.

Source	Articles	Source	Articles
freedomoutpost.com	1,638	personalliberty.com	434
frontpagemag.com	1,259	remnantnewspaper.com	139
lewrockwell.com	821	thewashingtonstandard.com	115
shtfplan.com	778	breaking911.com	73
vdare.com	468	clashdaily.com	12

Table 7
Statistics about the QProp corpus including the number of articles and their average length (tokens).

Label	Sources	Articles	Train	Dev	Test	Length (tokens)
Propagandistic	10	5,737	4,021	575	1,141	1084.46 ± 890.59
Non-propagandistic	94	45,557	31,972	4,564	9,021	620.31 ± 518.92
Total	**104**	**51,294**	**35,993**	**5,139**	**10,162**	672.22 ± 590.98

Entropy only.

We used two basic evaluation measures: F_1-measure and accuracy. For the multi-class setting in experiment 1, we report macro-averaged F_1, while for the binary setting in experiments 2 and 3, we take propaganda as the positive class and we compute F_1 with respect to that class (no macro-averaging). In order to better analyze the results, we used the McNemar statistical test (Japkowicz & Shah, 2011, p. 226). This is a non-parametric test that computes statistics based on the comparison between the number of instances in which the predictions of two classifiers differ. Such statistics approximate a χ^2 distribution, assuming that the number of instances in which the two predictors differ is greater than 20, a condition which we checked was always satisfied in our experiments. We selected the standard value of $\alpha = 0.05$. Therefore, whenever we use the term *statistically significant*, we refer to McNemar's test at 95% confidence level.

6.1. Experiment 1: four-way classification on the TSHP-17 corpus

Whereas identifying propagandistic articles is our main objective, here we replicate (Rashkin et al., 2017). Thus, we use a Maximum Entropy classifier to discriminate between the four classes in the TSHP-17 corpus: *trusted, hoax, satire*, and *propaganda*. Rashkin et al. (2017) relied on *word n-gram features* only (cf. Section 4.1). We also use these representations for this and the other experiments, and we consider them as a baseline. Our results using word *n*-grams on the original in- and out-of-domain partitions of the TSHP-17 corpus —including the void instances we discard for the rest of the experiments (cf. Section 5.1)— are $F_1 = 93.77$ and 66.99, respectively. These are slightly higher than the results reported in Rashkin et al. (2017) (91 and 65, respectively), but we consider the model to have been successfully replicated. The evaluation results on the filtered corpora are slightly higher.

We performed an ablation study: using (*i*) each feature family in isolation and (*ii*) all but one. We study the performance of the resulting multi-class models when testing on articles from seen (in-domain) vs. unseen (out-of-domain) sources. The left-hand side of Table 8 shows the results for feature families in isolation. The first row corresponds to the baseline word *n*-grams. One important aspect, which was present in Rashkin et al. (2017) as well, is the huge performance drop of about 25 points absolute from in-domain to out-of-domain. The remaining feature families exhibit a similar behavior: the model is much worse when dealing with articles from unseen sources. The gaps are smaller, as in the 11-points difference when using lexicon features. Still, the performance for all feature families is below the baseline in the out-of-domain setting.

The right-hand side of Table 8 shows the results obtained with all but one family. The trend between in- and out-of-domain still

Table 8
Macro-averaged F_1 and accuracy when predicting *propaganda, hoax, satire*, or *trusted* on the TSHP-17 corpus. *In-domain* refers to testing on documents from the same sources as in training, while *out-of-domain* means testing on documents from sources unseen on training.

only with	in-domain		out-of-domain		all but	in-domain		out-of-domain	
	F_1	Acc	F_1	Acc		F_1	Acc	F_1	Acc
word *n*-grams	94.46	94.41	**69.18**	**69.67**	word *n*-grams	**97.58**	**97.58**	67.13	67.77
lexicon	59.17	59.92	47.48	48.00	lexicon	96.77	96.76	65.32	66.57
voc. richness	47.63	50.25	31.35	34.90	voc. richness	96.82	96.81	65.65	66.83
readability	35.17	39.43	26.19	29.13	readability	96.77	96.76	65.47	66.70
char *n*-grams	**97.15**	**97.14**	63.98	65.00	char *n*-grams	95.43	95.38	**69.56**	**70.53**
nela	87.07	87.06	63.81	65.33	nela	96.89	96.89	63.96	65.27
					all	96.77	96.76	65.35	66.60

Table 9
Results on the two-way classification task (F_1): *propaganda* vs. *non-propaganda* on the in-domain partition of the TSHP-17 and on both partitions of QProp.

	TSHP-17	QProp	
Features	in-domain	Dev	Test
word *n*-grams	90.76	74.42	75.55
lexicon	68.74	46.55	44.87
voc. richness	55.62	29.45	29.72
readability	40.16	21.96	21.50
char *n*-grams	**96.22**	**82.93**	**82.13**
nela	82.27	54.60	50.98
word *n*-grams + char *n*-grams	**97.21**	78.37	79.01
char *n*-grams + lexicon	97.14	83.02	81.94
char *n*-grams + nela	96.64	**83.21**	82.75
readability + nela	82.30	75.34	76.83
char *n*-grams + lexicon + voc. richness + nela	96.97	83.17	**82.89**
word & char *n*-grams + lexicon + voc. richness + nela	97.10	79.04	79.50

holds. The best combination of features on the in-domain partition is precisely the one that excludes the word *n*-grams; it is even better than when considering all representations together. A similar phenomenon occurs on the out-of-domain case, but when excluding the character *n*-grams. These results suggest that all representations provide information that reflects the source rather than the four target classes. Therefore, we need to (*i*) train on corpora with a wider variety of sources in order to avoid the classifier to get confused and to learn source-specific patterns and (*ii*) do it while addressing our actual task: discriminating propaganda from non-propaganda. This leads to our experiment 2.

6.2. Experiment 2: two-way classification on TSHP-17 and QProp

Since we are interested in the binary task of distinguishing *propaganda* vs. *non-propaganda*, we asked ourselves whether the same drop between in-domain and out-of-domain articles manifests in the binary classification setting as well. We perform our analysis on both corpora (cf. Section 5). For the TSHP-17 corpus, we do one vs. the rest by converting *trusted, hoax* and *satire* articles into the negative class and we test on the in-domain partition only. QProp is already a two-way classification corpus.

Table 9 shows the results. The corpora are highly imbalanced now, and thus we will not show accuracy values. We first focus on the TSHP-17 corpus. The baseline word *n*-grams hold their status as a simple yet powerful representation, achieving an F_1 of 90.76. Nevertheless, whereas the other representations show a performance from average to poor, one representation stands out: character *n*-grams yield an F_1 of 96.22 (+5.46 with respect to word *n*-grams). The results on the QProp corpus, shown on the right-hand side of Table 9, follow the same trend. Once again, word and character *n*-grams perform better than the remaining representations.

On a corpus with ten propagandistic sources, character *n*-grams outperform word *n*-grams by five or more points in both partitions —82.93 (+8.51) and 82.13 (+6.58). These differences between the word and character *n*-gram are statistically significant.

The bottom part of Table 9 shows the results on both corpora with different combinations of feature families.[20] The feature combination improves the performance significantly, i.e. in most cases the different feature families capture different aspects. On the TSHP-17 corpus, combining word and character *n*-grams boosts the performance by one point absolute with respect to the model using character *n*-grams only. The results on the development and on the test partitions of QProp vary: the best combination on development is character *n*-grams and NELA, whereas adding lexicon and vocabulary richness on top of them works best on test. Nevertheless, the difference between the results with this combination and the character *n*-grams alone is not statistically significant.

[20] We explored all combinations, but here we only report a subset of the most interesting results. The rest are available at http://proppy.qcri.org.

Table 10
Results on the two-way classification task on the TSHP-17 corpus (F_1): *propaganda* vs. *non-propaganda* when testing on unseen sources.

Features	TSHP-17 corpus out-of-domain
word *n*-grams	50.68
lexicon	61.54
voc. richness	54.29
readability	45.68
char *n*-grams	52.51
nela	64.00
word *n*-grams + char *n*-grams	63.66
char *n*-grams + lexicon	52.89
char *n*-grams + NELA	53.66
readability + NELA	**64.14**
char *n*-grams + lexical + voc. richness + NELA	63.47

Table 11
Statistics of the re-distribution of the QProp corpus, created for experiment 3.

Class	Train		Test	
	Sources	Articles	Sources	Articles
Propagandist	5	2,802	5	2,935
Non-propagandist	47	22,776	47	22,781
Total	**52**	**25,578**	**52**	**25,716**

Although we do not select the types of features on a separate dataset and we only perform an *a posteriori* analysis of the results, we notice that most combinations yield comparable results. Thus, we can say that by picking any of these the results would not change significantly. In order to put these results into perspective, we computed a feature representation using ELMo embeddings (Peters et al., 2018). We used the pre-trained embeddings and we fed the biLM with (*i*) the title of the article, and (*ii*) the title and the first two sentences of the article. In both cases, we extracted the resulting representation: one 1024-dimensional vector per article. The F_1 score on the QProp corpus on the dev and test sets using the title only are 67.16 and 66.71, respectively, while using the title and the first two sentences of the article yielded 68.17 and 64.52, respectively.

After analyzing the results in Table 9, one question remains open: is the classifier still learning the sources rather than propaganda on the QProp corpus? In order to address this question, we perform experiment 3.

6.3. Experiment 3: learning propaganda vs. learning the source

In this experiment, we aim at analyzing whether our models learn to distinguish propagandistic vs. non-propagandistic articles as opposed to learning to recognize the news source an article is coming from. In order to do that, we first evaluate our models trained on the TSHP-17 corpus on its out-of-domain partition; i.e. on articles from unseen sources. Table 10 shows the results. The *char n-grams* ($F_1 = 52.51$), vocabulary richness ($F_1 = 54.29$), and *NELA* ($F_1 = 64.00$) features clearly improve with respect to the *word n-grams* ($F_1 = 50.68$), and the improvements are statistically significant.

The information available in our QProp corpus regarding the source of each article (cf. Section 5.2) allows for a more sophisticated experiment. In particular, we reshape QProp by performing the following steps: (*i*) we merge the training, the development, and the testing partitions into one single collection; (*ii*) we randomly split the positive (negative) instances into two subsets: Qprop_1^+ and Qprop_2^+ (Qprop_1^- and Qprop_2^-); and (*iii*) we compose a new training set by mixing Qprop_1^+ and Qprop_1^- and a new testing set by mixing Qprop_2^+ and Qprop_2^-. We apply a number of constrains when producing this redistribution. First, we make sure there is no intersection between the sources in the new training and testing partitions. Second, we include an equal number of propagandistic and non-propagandistic sources in each partition. Third, we force the two propagandistic sources with less than 100 instances to be part of the test set (cf. Table 6). We perform several random samplings in order to come out with partitions as balanced as possible. Table 11 shows statistics about this version of the corpus.

We perform a number of experiments with an increasing number of instances on the training side, sampling subsets of positive instances according to their source. The procedure is as follows. Let $s_{1,\ldots,5}$ be the five propagandistic sources in the training set D_{tr}. We select at random $k \leq 5$ propagandistic sources and we keep only those documents belonging to the selected sources, resulting in D_{tr}^*. The negative instances are sub-sampled as well in order to resemble the distribution of the data in the original QProp, but regardless of their sources. We then train a model on the resulting D_{tr}^* and we evaluate it on the testing partition. We keep the test set untouched in all cases as, regardless of the sub-sampling, the models are always tested on articles whose sources, both propagandistic and non-propagandistic, were not seen during training. We repeated this experiment with all possible combinations of $k \in [1, 5]$ propagandistic sources and with all feature families.

Table 12
Results on the binary classification task (average F_1) in which articles from a subset of propagandistic and non-propagandistic sources are used for training and predictions are made only on articles from different sources. An increasing number of propagandistic sources, from 1 to 5, is used (note that there is no standard deviation when having 5 sources, as there is only one possible combination).

src	only word *n*-grams	only char *n*-grams	only nela	all features	all *n*-grams	all features but word *n*-grams	all features but char *n*-grams
1	2.78 ± 5.00	35.01 ± 17.88	**38.91 ± 11.92**	10.27 ± 10.49	39.78 ± 12.87	35.89 ± 18.60	7.16 ± 8.60
2	11.36 ± 9.57	50.20 ± 8.37	44.71 ± 1.85	26.05 ± 10.76	45.88 ± 2.05	**52.01 ± 8.39**	20.24 ± 10.07
3	23.24 ± 9.77	57.73 ± 3.86	45.93 ± 1.12	38.35 ± 7.21	47.15 ± 1.20	**59.48 ± 3.67**	32.30 ± 7.44
4	34.83 ± 6.89	62.04 ± 2.23	46.60 ± 0.38	46.79 ± 4.64	47.88 ± 0.47	**63.51 ± 2.06**	41.94 ± 5.07
5	44.75	64.45	47.01	53.36	48.25	**65.61**	49.95

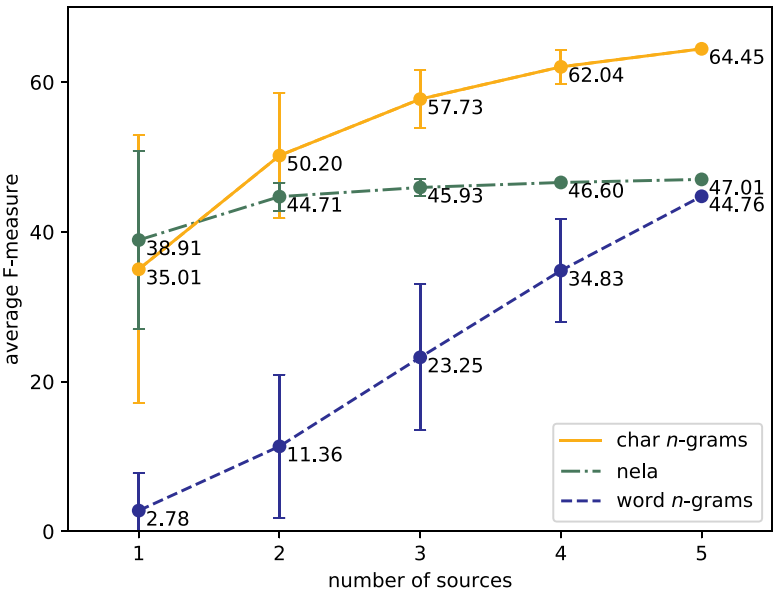


Fig. 2. Zooming into the results on the binary classification task (average F_1) for experiment 3 when using individual feature families (cf. Table 12 for further details).

Table 12 shows a selection of the results.[21] Note that these results are not comparable with the ones from Table 9 since the training and test sets are different. In the table, we report means and standard deviation. For the experiments with one or two sources only, the standard deviation values are high. This might reflect the high variability in the number of articles per source (cf. Table 6), which translate into relatively small and widely different training set sizes.

As it happened for the experiments on the TSHP-17 corpus, *char n-grams* and *nela* features perform significantly better than the *word n-grams*: the difference in terms of F_1 between the *char n-gram* and the *word n-gram* features ranges from 38.84 (when training on two sources) to 19.70 (when training on five sources); the difference in F_1 between the *nela* and the *word n-gram* features ranges from 36.13 (when training on two sources) to just 2.26 when training on five sources. Fig. 2 zooms into these three experiments with single feature families to give a more clear picture. Whereas the *nela* features perform the best on average when learning from one source only, they are not so good when training on more sources, arriving close to convergence with the word *n*-grams when considering five sources (47.01 vs. 44.75). Overall, character *n*-grams are the best individual type of features. Even if they perform slightly worse than the *nela* features when learning from one positive source only, they significantly outperform the others as soon as they have access to positive instances from two or more sources, reaching a difference of 17.44 absolute over *nela* when given access to five sources. This experiment clearly shows that *word n-grams* are not effective when testing on articles from unseen sources.

The performance evolution when considering all the features, shown in the middle column of Table 12, suggests that considering the word *n*-grams causes the classifier to under-perform; this is also the case when using these types of features alone. In order to confirm this, we performed an ablation study where we excluded the character *n*-grams, the word *n*-grams, and also both. The rightmost three columns of Table 12 show the results. We can see that when excluding both the word and the character *n*-grams, the classifier is dominated by the *nela* features, which are the best performing ones among those considered. Yet, the model behaves as we expected when excluding either the word or the character *n*-grams. In the former case, the character *n*-grams complement well the rest of the features, and yield a slight boost in performance —yielding the best results overall— after training on at least two sources.

[21] The full set of results is available at http://proppy.qcri.org/about.html

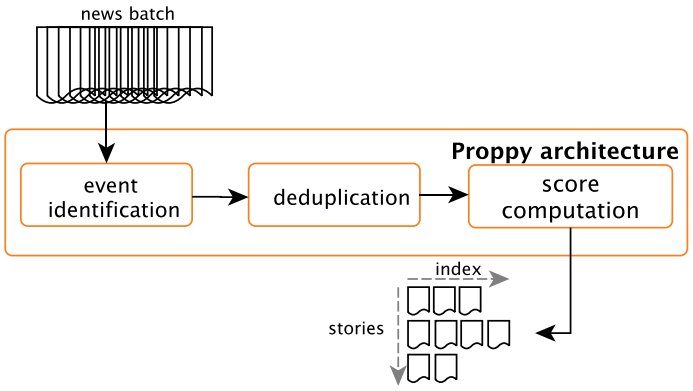


Fig. 3. Proppy's architecture and modules: event identification, deduplication, and score computation.

In the latter case, excluding the character *n*-grams makes the classifier worse and the negative effect of the word *n*-grams is much higher.

Overall, the outcome of our three experiments shows that the models that use representations modeling writing style and readability always outperform those based on word-level representations. This is particularly true in those cases where the training and the testing datasets do not include articles from the same news outlets. As a result, we can say that Hypothesis 1 has been confirmed. Thus, we can conclude that when writing and detecting propaganda, style matters more than topic.

7. Prototype architecture

We further developed a prototype to demonstrate our propaganda identification model in action.[22] Fig. 3 shows an overview of its architecture. The process begins when a batch of news articles is fed to the system. We rely on GDELT to retrieve articles, as for the construction of the `QProp` corpus (cf. Section 5.2), but this time live: we process articles from 56 sources every 24 h. The first module is in charge of identifying events in the batch of articles. As in `Qlusty` (Barrón-Cedeño, Da San Martino, Zhang, Ali, & Dalvi, 2018), we perform a DBSCAN clustering (Ester, Kriegel, Sander, & Xu, 1996) using doc2vec representations (Le & Mikolov, 2014) of the news articles. The second module discards near-duplicates. Once again, as Barrón-Cedeño et al. (2018), we compute the Jaccard coefficient (Jaccard, 1901) between all pairs of articles in an event and, if the result surpasses a given threshold, one of them is discarded. The resulting set of articles is assessed for propaganda.

In our prototype, we opt not to make a binary decision in absolute terms, i.e. whether a given document is propagandistic or not. This is because an entire news article could hardly be flagged as entirely propagandistic. It could just contain pieces or propaganda; this has been observed in the case of fake vs. real news as well (Potthast et al., 2017). Instead, we opt for estimating a *propaganda score*. This score intends to reflect to what extent a piece of news might have a propagandistic intent. The score is calculated by our binary prediction model and it lies in the range [0,1]. The output is a matrix of stories and propaganda scores.

Once the user has selected an event, the articles it covers get organized into five bins, according to the different levels of propagandistic intent that our model has detected in their contents. Fig. 4 shows two snapshots of the interface. Fig. 4a shows two articles covering the Trump–Putin reunion in July 2018.[23][24] Our model estimates propaganda scores of 0.282 and 0.323 and locates them in the second bin. That is, they have a relatively low level of propagandistic intent. Fig. 4b shows an opinion article discussing the aftermath of the same reunion.[25] Here, our model estimates a higher propaganda score of 0.95 and locates it in the fifth bin, which corresponds to the highest propagandistic intent. In this way, the user can observe how different media talk about the same event on the propaganda dimension, which may guide her in her further exploration, even considering bias and factuality.

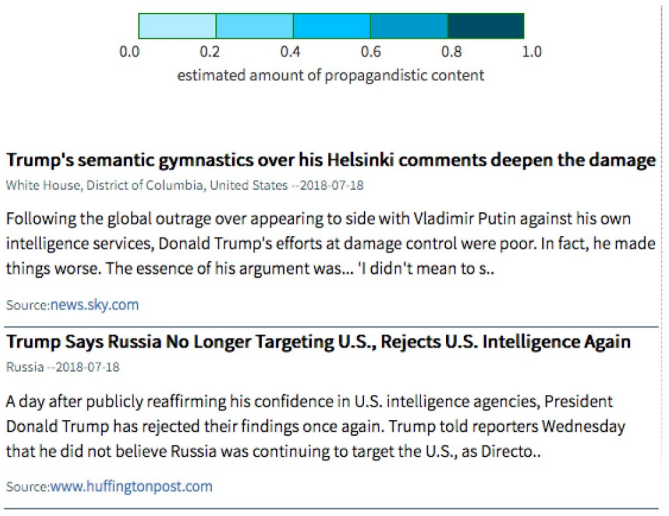
This architecture follows a *push* publishing model. That is, it is the system that updates automatically the material that it presents to the user without her taking any action but exploring the available events. Fig. 5 shows an alternative architecture, which follows a *pull* model. In this case, the user queries the system with a topic of her interest (e.g., a character, an event). The search engine ranks the documents with a standard scoring function such as BM25 (Manning, Raghavan, and Schütze, 2008, p. 232) and, once again, the score computation model estimates a propaganda score. The relevant articles can then be displayed according to a combination of the relevance against the query and time on one axis and according to the propaganda score on the other.

[22] The prototype is available at http://proppy.qcri.org.

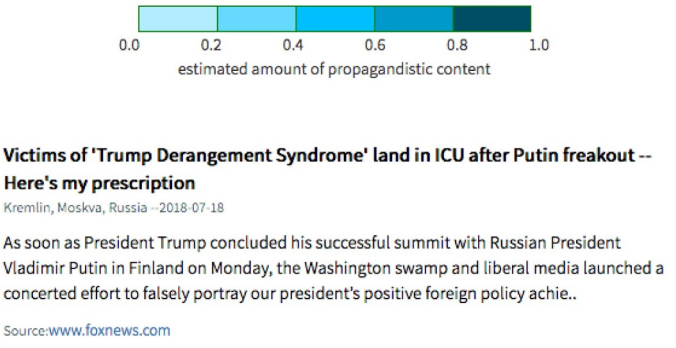
[23] https://news.sky.com/story/trumps-semantic-gymnastics-over-his-helsinki-comments-deepen-the-damage-11440277

[24] https://www.huffingtonpost.com/entry/donald-trump-russia-cyberattacks_us_5b4f6ec9e4b0fd5c73c16577

[25] http://www.foxnews.com/opinion/2018/07/18/victims-trump-derangement-syndrome-land-in-icu-after-putin-freakout-heres-my-prescription.html.



(a) Two articles from the second propaganda score bin.



(b) One article from the fifth propaganda score bin.

Fig. 4. Snapshots of the prototype showing articles covering (a) the Trump–Putin meeting in Helsinki in July 2018 and (b) an opinion column analyzing the aftermath.

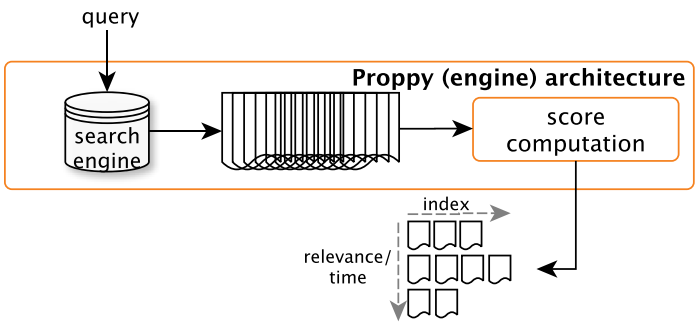


Fig. 5. An alternative architecture, where the user queries a search engine, and the relevant articles are organized according to their relevance and propaganda score.

8. Conclusion and future work

We performed a thorough experimentation into propaganda detection at the news article level. Our experimental results show that representations modeling writing style and text complexity are more effective than word *n*-grams, which model topics. Our comparison against existing models corroborates this hypothesis: models that consider stylistic features, such as character *n*-grams always outperform alternative representations, which are typically used in topic-related tasks. Different from previous approaches,

this is true also when trying to classify articles from sources unseen on training. This is a key asset when dealing with the never-ending spawn of news outlets: propagandistic vs. other.

We further presented a system that organizes news articles into events and, for each event, shows articles according to their level of propagandistic content. The system is designed with the aim of raising awareness into individual readers as well as providing tools for organizations to monitor large amounts of news articles. Finally, we published an interface where our system organizes events according to propaganda, we also released the source code used in these experiments as well as our new corpus. We believe that these three resources are valuable for further research on propaganda detection, and that they will be also appreciated by the research community as well as by the general public.

Interesting avenues for future research include going into the fragment level and training models to identify specific propaganda techniques. That would allow for the creation of models able to explain their decisions and to give the user a clearer picture of what propagandistic techniques have been put in use.

Appendix A. Analysis of the most relevant word *n*-grams

In this appendix, we look at the most informative word *n*-grams as considered by the classifier to differentiate between propagandistic and non-propagandistic articles. In order to do that, we built a binary classifier on the `QProp` corpus only with word *n*-grams and we retrieved the strings that the model assigned the highest weights to —both for the propaganda and for the non-propaganda classes.

Tables A.13 and A.14 show the most important word *n*-grams that help the classifier to decide whether a text should be classified as propagandistic or not. As Table A.13 shows, strings that refer to posting a piece of news after getting proper permission from another source (block 1) are among those with the highest weights. This may reflect that propagandistic articles tend to be re-posted in different media. Other strings are more related to superlatives (e.g., block 2). Also, three blocks include strings associated with people profiling (blocks 4, 6, 8; perhaps also related to the so-called *flag waving* and *bangwagon* propagandistic techniques), or to specific characters (block 7). It is worth noting that the characters mentioned in block 7 have less media presence nowadays; therefore, relying on them to identify propaganda is a time-sensitive issue.

On the other extreme, Table A.14 shows the highest-weighted word *n*-grams for the negative class —non-propaganda. It is interesting to note that strings with “said” and related verbs (block 1) are among those with the highest weights. This might reflect than non-propagandistic articles tend to quote the actors or reporters of the events. Having most weekdays reflects something similar (block 3): it is more likely that non-propagandistic news will cover a punctual event occurring at a specific time, rather than columns and other kinds of pieces. Tables A.15 and A.16 show some instances of these (groups of) strings in context. This small subset of examples shows that indeed the *n*-grams associated with propagandistic articles tend to occur in propagandistic text snippets, whereas those associated with non-propaganda tend to occur in more neutral and objective sentences.

Table A.13
Top-18 most significant word *n*-grams for the propaganda class (stopword instances not shown); **b** = block of (semantically)-related instances (it links to the examples in Table A.15), **w** = weight assigned by the classifier.

b	w	*n*-gram	b	w	*n*-gram
1.	1.17	with permission	4.	0.49	american
	1.09	permission from	5.	0.47	the left
	0.99	article posted	6.	0.46	muslim
	0.98	article posted with	7.	0.46	obama
	0.97	posted with permission		0.45	clinton
	0.95	posted with	1.	0.43	originally published by
2.	0.63	the best of		0.43	whole article
	0.62	best of	8.	0.43	united states
3.	0.52	actually	1.	0.43	the whole article

Table A.14
Top-18 most significant word *n*-grams for the negative class (instances with stopwords not shown); **b** = block of (semantically)-related instances (it links to the examples in Table A.16), **w** = weight of the classifier.

b	w	*n*-gram	b	w	*n*-gram	b	w	*n*-gram
1.	−1.69	said	1.	−0.39	told	3.	−0.33	saturday
	−0.63	said the	4.	−0.38	minister		−0.31	week
2.	−0.47	after	3.	−0.35	wednesday		−0.31	monday
1.	−0.44	he said		−0.34	tuesday	5.	−0.31	photo
2.	−0.43	last	1.	−0.33	said in	6.	−0.29	provided
3.	−0.41	thursday	3.	−0.33	friday	7.	−0.29	read more

Table A.15
Collocation-like examples including the word *n*-grams in Table A.13 (linked by the number on the left of each block).

1.	Article	**posted with permission**	from Robert Spencer
	party or has been republished	**with permission**	from the author
	Article	**posted with permission**	from End of the American Dream
	This article was	**originally published**	by Adam Taggart at PeakProsperity.com
	This report was	**originally published**	by Jeremiah Johnson at Tess Pennington's
2.	their anger don't represent	**the best of**	America, they represent the worst of
	and gave us	**the best of**	medieval law
	after fellow venture members to	**the best of**	his ability
3.	If the NRA	**actually**	cared about the Second Amendment
	thereby endangering them, when	**actually**	all I did was respond to published
	the family noted that Roberson was	**actually**	wearing security attire
4.	Speaking at an African	**American**	church in Boynton Beach
	disgraceful in all of this is that the	**American**	people were promised a special prosecutor
	the increasing balkanization of the	**American**	body politic
5.	now expressing hatred (which	**the left**	does so well) rather than love
	the greatest existential threat	**the left**	has ever faced in America
	How has	**the left**	elite handled these allegations?
6.	the more	**muslim**	savages we allow into america
	There is no such thing as a moderate	**muslim**	and there never will be
	Ally will be the first	**muslim**	male Judge in New York
7.	Barack Hussein	**Obama**	Soetoro Sobarkah
	Americans praised him under	**Obama**	and demonize him under Trump
	Why aren't they going after Hillary	**Clinton**	with her emails and with the dossier
8.	If that actually happened in the	**United States**	of America and everything each and every
	the Missile Defense Agency and the	**United States**	government in their ballistic missile defense
	this "sticks in the craw" of the	**United States**	and the Western Financial,

Table A.16
Collocation-like examples including the word *n*-grams in Table A.14 (linked by the number on the left of each block).

1.	With that	**said**	, while President Donald Trump
	Republican Congressman Trey Gowdy	**said**	he thought it was politically smart
	Tempe Police Department,	**said**	the women were arrested
2.	video footage released	**after**	the official meeting
	As TFTP reported	**last**	week, Carol Davidsen
	This monstrous slaughter took place	**last**	October, and still the FBI has nothing
3.	Miami collapsed on	**Thursday**	, possibly killing several motorists
	shooter drills at the school that very	**week**	and that they would be firing
	President Trump on	**Wednesday**	voiced support for confiscating guns
4.	A Lutheran	**minister**	and early Nazi supporter
	Does the	**Minister**	agree that Tommy Robinson
	by Home Office	**Minister**	Ben Wallace
5.	that he posted a	**photo**	on Facebook
	Relying on a	**photo**	posted on Collins Facebook
	Then after the	**photo**	was taken
6.	tested for a rape kit and she	**provided**	a written account
	Brandon Curtis at Concealed Nation	**provided**	some thoughts
	was not based on any information	**provided**	to her by Obama himself
7.		**Read more**	about the Thursday activities here
		Read more	about that by clicking linked
	document here, and	**read more**	about it here

References

Ba, M. L., Berti-Equille, L., Shah, K., & Hammady, H. M. (2016). *VERA: A platform for veracity estimation over web data. Proceedings of the 25th International Conference Companion on World Wide WebWWW '16*159–162 Montréal, Québec, Canada

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM, 61*(6), 54–61.

Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., & Nakov, P. (2018). *Predicting factuality of reporting and bias of news media sources. Proceedings of the 2018 Conference on Empirical Methods in nAtural Language ProcessingEMNLP '18*3528–3539.

Barrón-Cedeño, A., Da San Martino, G., Zhang, Y., Ali, A., & Dalvi, F. (2018). *Qlusty: Quick and dirty generation of event videos from written media coverage. Proceedings of the Second International Workshop on Recent Trends in News Information Retrieval*27–32 Grenoble, France

Bazerman, C. (2010). *The informed writer: Using sources in the disciplines.* Fort Collins, CO, USA: The WAC Clearinghouse.

Bird, S., Loper, E., & Klein, E. (2009). *Natural language processing with python.* O'Reilly Media Inc.

Brill, A. M. (2001). Online journalists embrace new marketing function. *Newspaper Research Journal, 22*(2), 28.

Canini, K. R., Suh, B., & Pirolli, P. L. (2011). *Finding credible information sources in social networks based on content and social structure. Proceedings of the IEEE International Conference on Privacy, Security, Risk, and Trust, and the IEEE International Conference on Social ComputingSocialCom/PASSAT '11*1–8 Boston, Massachusetts, USA

Castillo, C., Mendoza, M., & Poblete, B. (2011). *Information credibility on Twitter. Proceedings of the 20th International Conference on World Wide WebWWWZ'11*675–684 Hyderabad, India
Chen, C., Wu, K., Srinivasan, V., & Zhang, X. (2013). *Battling the Internet Water Army: Detection of hidden paid posters. Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and MiningASONAM '13*116–120 Niagara, Ontario, Canada
Conserva, H. (2003). *Propaganda techniques.* United States: AuthorHouse.
Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press.
Davies, W. (2016). *The age of post-truth politics.* The New York Times.
Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G., & Zubiaga, A. (2017). *SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. Proceedings of the 11th International Workshop on Semantic EvaluationSemEval '17*60–67 Vancouver, Canada
Ellul, J. (1965). *Propaganda: The formation of men's attitudes.* United States: Vintage Books.
Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data MiningKDD '96*226–231 Portland, OR, USA
Finberg, H., Stone, M. L., & Lynch, D. (2002). *Digital journalism credibility study. 3*, Online News Association. Retrieved November2003.
Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029.
Gunning, R. (1968). *The technique of clear writing.* McGraw-Hill.
Hardalov, M., Koychev, I., & Nakov, P. (2016). *In search of credible news. Proceedings of the 17th International Conference on Artificial Intelligence: Methodology, systems, and applicationsAIMSA '16*172–180 Varna, Bulgaria
Honore, A. (1979). Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin, 7*(2), 172–177.
Hooper, J. (1975). *On assertive predicates. 4*. Academic Press, New York.
Horne, B., Khedr, S., & Adal, S. (2018). *Sampling the news producers: A large news and feature data set for the study of the complex media landscape. Proceedings of the 12th International AAAI Conference on Web and Social MediaICWSM '18*518–527 Stanford, CA, USA
Horne, B. D., & Adal, S. (2017). *This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. Proceedings of the international workshop on news and public opinion at ICWSM* Montreal, Canada
Hyland, K. (2015). *The international encyclopedia of language and social interaction. The International Encyclopedia of Language and Social Interaction.* American Cancer Society1–11).
How to Detect Propaganda (1938). In Institute for Propaganda Analysis (Ed.). *Propaganda analysis. volume i of the publications of the institute for propaganda analysis* (pp. 210–218). New York, NY
Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles, 37*, 547–579.
Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: A classification perspective.* New York, NY: Cambridge University Press.
Jowett, G., & O'Donnell, V. (2012). *Propaganda and persuasion* (5th). Los Angeles, CA, USA: SAGE.
Juola, P. (2012). An overview of the traditional authorship attribution subtask. In P. Forner, J. Karlgren, & C. Womser-Hacker (Eds.). *CLEF 2012 evaluation labs and workshop – working notes papers* Rome, Italy
Kincaid, J., Fishburne, R., Rogers, R., & Chissom, B. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnelTechnical Report 1-1-1975*. Springfield, Virginia: Memphis TN Naval Air Station.
Koppel, M., Schler, J., & Bonchek-Dokow, E. (2007). Measuring differentiability: Unmasking pseudonymous. *Journal of Machine Learning Research, 8*, 1261–1276.
Kulkarni, V., Ye, J., Skiena, S., & Wang, W. Y. (2018). *Multi-view models for political ideology detection of news articles. Proceedings of the 2018 Conference on Empirical Methods in Natural Language ProcessingEMNLP '18*3518–3527 Brussels, Belgium
Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., et al. (2018). The science of fake news. *Science, 359*(6380), 1094–1096.
Le, Q., & Mikolov, T. (2014). *Distributed representations of sentences and documents. Proceedings of the 31st international conference on international conference on machine learning - volume 32ICML '14* I(I–1188–II–1196). Beijing, China
Leetaru, K., & Schrodt, P. A. (2013). *GDELT: Global data on events, location, and tone, 1979–2012. International Studies Association Meetings2. International Studies Association Meetings* 1–49.
Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., et al. (2016). A survey on truth discovery. *SIGKDD Explorations Newsletter, 17*(2), 1–16.
Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval.* New York, NY, USA: Cambridge University Press.
Mihaylov, T., Georgiev, G., & Nakov, P. (2015). *Finding opinion manipulation trolls in news community forums. Proceedings of the Nineteenth Conference on Computational Natural Language LearningCoNLL '15*310–314 Beijing, China
Mihaylov, T., Mihaylova, T., Nakov, P., Màrquez, L., Georgiev, G., & Koychev, I. (2018). The dark side of news community forums: Opinion manipulation trolls. *Internet Research, 28*(5), 1292–1312.
Mihaylova, T., Karadjov, G., Pepa, A., Baly, R., Mohtarami, M., Barrón-Cedeño, A., et al. (2019). *SemEval-2019 task 8: Fact checking in community question answering forums. Proceedings of the international workshop on semantic evaluationSemEval '19* Minneapolis, MN, USA
Muller, R. (2018). Indictment of Internet Research Agency.
Nakov, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., et al. (2018). *CLEF-2018 lab on automatic identification and verification of claims in political debates. Working notes of CLEF 2018 – Conference and Labs of the Evaluation ForumCLEF '18*372–387 Avignon, France
Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *LIWC Operators Manual 2001*.
Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations. Proceedings of the 2018 Conference of the North American chapter of the Association for Computational Linguistics: Human language technologiesNAACL-HLT '18*2227–2237 New Orleans, LA, USA
Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2017). *Where the truth lies: Explaining the credibility of emerging claims on the Web and social media. Proceedings of the 26th International Conference on World Wide Web CompanionWWW '17 Companion*1003–1012 Perth, Australia
Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2017). A stylometric inquiry into hyperpartisan and fake news. *CoRR, abs/1702.05638*.
Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). *A stylometric inquiry into hyperpartisan and fake news. Proceedings of the 56th annual meeting of the Association for Computational LinguisticsACL '18*231–240 Melbourne, Australia
Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). *Truth of varying shades: Analyzing language in fake news and political fact-checking. Proceedings of the 2017 Conference on Empirical Methods in Natural Language ProcessingEMNLP '17*2931–2937 Copenhagen, Denmark
Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *SIGKDD Explorations Newsletters, 19*(1), 22–36.
Sproule, J. M. (2001). Authorship and origins of the seven propaganda devices: A research note. *Rhetoric & Public Affairs,, 4*(1), 135–143.
Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*(3), 538–556.
Thorne, J., & Vlachos, A. (2018). *Automated fact checking: Task formulations, methods and future directions. Proceedings of the 27th International Conference on Computational LinguisticsCOLING '18*3346–3359 Santa Fe, NM, USA
Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). *FEVER: A large-scale dataset for fact extraction and VERification. Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologiesNAACL-HLT '18*809–819 New Orleans, LA, USA
Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.
Wilson, T., Wiebe, J., & Hoffmann, P. (2005). *Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language ProcessingNAACL-HLT '05*347–354 Vancouver, Canada
Yule, G. (1944). *The statistical study of literary vocabulary.* Cambridge University Press.
Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE, 11*(3), 1–29.