HAL
open science

# Sifting the Arguments in Fake News to Boost a Disinformation Analysis Tool

Jérôme Delobelle, Amaury Delamaire, Elena Cabrio, Ramón Ruti, Serena Villata

## To cite this version:

Jérôme Delobelle, Amaury Delamaire, Elena Cabrio, Ramón Ruti, Serena Villata. Sifting the Arguments in Fake News to Boost a Disinformation Analysis Tool. NL4AI 2020 - 4th Workshop on Natural Language for Artificial Intelligence, Nov 2020, Online, Italy. hal-02990781

**HAL Id: hal-02990781**

**https://hal.science/hal-02990781v1**

Submitted on 5 Nov 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Sifting the Arguments in Fake News to Boost a Disinformation Analysis Tool

Jérôme DELOBELLE[1], Amaury DELAMAIRE[2], Elena CABRIO[1], Ramón RUTI[2], and Serena VILLATA[1]

[1] Université Côte d'Azur, Inria, CNRS, I3S, Sophia-Antipolis, France
jerome.delobelle@u-paris.fr
{elena.cabrio,serena.villata}@unice.fr
[2] Storyzy, 130 rue de Lourmel 75015 Paris, France
{amaury.delamaire,ramon.ruti}@storyzy.com

**Abstract.** The problem of disinformation spread on the Web is receiving an increasing attention, given the potential danger fake news represents for our society. Several approaches have been proposed in the literature to fight fake news, depending on the media such fake news are concerned with, i.e., text, images, or videos. Considering textual fake news, many open problems arise to go beyond simple keywords extraction based approaches. In this paper, we present a concrete application scenario where a fake news detection system is empowered with an argument mining model, to highlight and aid the analysis of the arguments put forward to support or oppose a given target topic in articles containing fake information.

**Keywords:** Argument mining · Stance Detection · Fake news.

## 1 Introduction

The phenomenon of disinformation, produced and transmitted on the Web via social media platforms, websites, and forums is not new, but it has taken on an unprecedented scale since 2016 with the campaign for the American presidential election, the Brexit campaign, and in 2017 with the French presidential campaign. These days the health emergency associated with Covid-19 has exacerbated this problem considerably. If the phenomenon of disinformation has a very long history in the form of rumor, it has taken, in the digital age, first the name of "fake news" then that of "fake information". A fake news deals both with the dissemination of information without certainty whether it is false or true, but also with the intention pursued by the author of the content to mislead the audience.

Given the potential danger fake news represent for the users, several approaches have been developed to address the automatic detection of fake news online (e.g.,

*Corresponding author: jerome.delobelle@u-paris.fr*
Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

see [17]). The proposed methods may vary depending on the type of media the fake news is spread upon, i.e., text, images, or videos. Among those, we find algorithms that automatically identify fake news by recognising keywords in the text [9], by detecting the reuse of misleading images using their chronology, and by detecting the angle of the face, its expression, the lighting and other important information to verify the authenticity of videos about people [15]. In general, the task of fake news detection aims at easing the work of human analysts that have to investigate the ways fake news spread on the Web to then find a way to limit this phenomenon. Therefore, providing automatic (or semi-automatic) tools to make the analysis of fake news more effective to analysts is a main open challenge.

In this paper, we address this issue by proposing an argumentation-based disinformation analysis tool to support analysts to investigate the diffusion of fake news according to several criteria. More precisely, we propose to extend a disinformation analysis tool with a stance detection module [5] relying on pretrained language models, i.e., BERT [3], with the aim of obtaining a more effective analysis tool both for users and analysts. To evaluate the stance detection module in the disinformation context, we propose to annotate a new resource of fake news articles, where arguments are classified as being *InFavor* or *Against* towards a target topic. Our new annotated data set contains sentences about three topics currently attracting a lot of fake news around them, i.e., public health demands vaccination, white helmets provide essential services, and the risible impact of Covid-19. This data set collects 86 articles containing nearly 3000 sentences.

Although argumentation is an area of research in Artificial Intelligence that has received an increasing attention in recent years, few links have been made to bridge it with the detection of fake news. Among these works, Sethi [13] propose a prototypical social argumentation framework to verify the validity of proposed alternative facts to help in limiting the propagation of fake news. The debate about the veracity of a given fact is represented through a basic argumentative structure (claims, evidences and sources connected by a support relation and an attack relation), and it is crowdsourced and mediated by expert moderators in a virtual community. However, the argumentative part is limited to the formalisation of the debate, and no empirical evaluation is addressed. This approach does not deal with natural language arguments. More recently, Kotonya and Toni [4] introduce a new method for veracity prediction based on a form of argumentative aggregation. More precisely, they use stance label predictions for relation-based argument mining to generate a bipolar argumentation framework [2] which is a triple composed of a set of arguments, an attack and a support relation between arguments. Each argument is then evaluated with the DF-QuAD gradual semantics [11] in order to assess the veracity of the news against some evidence. However, their method applies to Twitter conversations, and more precisely, to the RumourEval dataset[3], where they consider each tweet in the conversation as a potential argument which discusses/disagrees/agrees with another tweet. They do not propose any concrete integration of their approach to a fake news detection system, as it is the case for our stance detection module.

[3] `http://alt.qcri.org/semeval2017/task8/`

In this paper, we do not present a new approach to detect fake news online. We propose an argumentation-empowered disinformation analysis tool to support analysts in a better understanding of the fake news content and structure, to conceive more effective solutions to fight its spread. To the best of our knowledge, this is the first concrete tool for fake news classification based on a stance detection module to aid the analysis of fake news and their diffusion.

The paper is organised as follows. Section 2 introduces the fake news detection module and its main features. In Section 3, we describe the manually annotated data set of fake news we created, and we report on the performances of the stance detection module on such dataset. Section 4 describes the resulting Disinformation Analysis Tool which combines the automated analysis of disinformation spread with the stance detection module. Conclusions end the paper, discussing directions for future work.

## 2 Fake News Detection

The disinformation analysis system we propose to extend in this work is the Storyzy[4] Disinformation Analysis tool (DAT), that automatically classifies information sources according to their reliability. In the following, we describe the main building blocks composing such system:

1. *Manual annotation of information sources according to their reliability.* Information sources are tagged by information experts and fact-checkers according to a restricted set of categories. More precisely, each source is annotated as belonging to one of these three categories: *trusted*, *fake news* or *satire*. For those sources that have been annotated as fake news, at least one more additional subcategory (e.g., *hate*, *conspiracy*, *propaganda*) is added.
2. *Numeric representation of information sources and training.* Information sources are turned into multidimensional vectors in order to build language models – one model for each information reliability category. Several language models can be used such as bag-of-words or ngrams, though ngrams are not easily scalable due to the size of the models. Different weighting functions are also available, such as tf-idf or frequency.
3. *Collection of new sources.* New information sources are gathered on a regular basis through diverse heuristics. These will feed the classifier and Storyzy's database.
4. *Classification of new sources.* New information sources spotted in step 3 are vectorised and classified according to the language models built in step 2. The classification is performed through probability computations which represent the chances of a new information source to belong to an information category.
5. *Verification of newly classified information sources.* The result of step 4 is analysed by human experts and added to the models from step 2. Thus the pipeline is circular and feeds itself with little human supervision. Human validation of newly classified information sources is based on several external criteria beyond classification probabilities. These criteria are mainly based on the analysis of information sources pointing at or pointed by the considered information source. The combination of

[4] `https://storyzy.com`

multiple criteria aims at reducing the workload of human experts by selecting the most relevant information sources.

It is worth noting that newly constructed language models are evaluated and compared to the previous ones in order to avoid regressions. While model evaluation is automated, the final decision of modifying a current model is left to human supervision. Human validations are always necessary in the delicate field of disinformation detection, but their load is kept to the minimum. A full automation of such a pipeline is not possible yet with satisfying results. This is due to the complexity of the task – even for humans – and the multiple engineering issues inherent to dynamic information monitoring.

The accuracy of Storyzy classification system is comparable to the human upper bound for this task with a score of 90% on a benchmark of more than 5500 English websites (where 1/4 are fake instances).[5] As the fake news detection system of Storyzy mainly relies on the assessment of the reliability of the news sources, we propose to extend such system with an argument mining module which instead focuses on the content of the news itself, with the aim to provide analysts with the arguments present in such articles and their stance toward the targeted topic, so that they can be supported in their decision to mark a news as being fake or not.

## 3 Stance Detection for Arguments in Fake News

Argument mining (AM) [7, 1, 6] is the research area aiming at extracting natural language arguments and their relations from text. The classic argument mining pipeline is composed of three main steps: first, the argument components are identified in the text; second, the boundaries of such components are defined; third, the intra-argument relations (relations among the evidences and the claims composing the same argument) and the inter-argument relations (relations among different arguments, e.g., support and attack) are predicted.

In this context, we focus on the specific task of *stance detection* which is commonly defined as the "automatic classification of the stance of the producer of a piece of text, towards a target, into one of these three classes: Favor, Against, Neither" [5]. The main rationale for this choice can be summarised as follows: given that the intent of the producers of disinformation is to destabilise populations from a political, economical and societal point of view, the stance of the fake news arguments with respect to the positions of established authorities is an important indicator to detect them, e.g., *"Covid-19 is not serious because the pandemic can be dramatically slowed, or stopped, with the immediate widespread use of high doses of vitamin C."*

Given our application scenario, i.e., stance detection for arguments from heterogeneous sources (newspaper articles, blogs, exchanges in online debate platforms) containing fake information and on different topics, we decided to adopt the model proposed by Stab et al. [14] which is both general and simple. They define an *argument* as a span of text expressing evidence or reasoning that can be used to either support or

[5] Further details both on the system architecture and on its performance cannot be disclosed due to copyright reasons.

oppose a given target topic. Furthermore, an argument may presuppose some domain knowledge, (or the application of commonsense reasoning), but it must be unambiguous in its orientation to the target topic. A *target topic*, in turn, is some matter of controversy for which there is an obvious polarity to the possible outcomes - that is, a question of being either in favor or against the use or adoption of something, the commitment to some course of action, etc. Thus, given a target topic, it is possible to label a sentence in one of the following three categories:

**(Argument-InFavor)** A sentence expressing evidence or reasoning that can be used to support a target topic.

**(Argument-Against)** A sentence expressing evidence or reasoning that can be used to oppose a target topic.

**(Neither)** This category includes all other sentences. In other words, it can be a non-subject-related argument or just a non-argumentative sentence that may or may not be related to the target topic (e.g., a definition).

Following this definition, in this work we address the task of topic-dependent, sentence-level stance detection. We cast it as a three-way classification task: given a sentence and a topic, the algorithm should classify it as either an *Argument-InFavor*, an *Argument-Against* or *Neither*. Given that the use of contextualized word embeddings is the approach offering the best results for this task [12], we opted for such method. Among existing approaches, BERT (Bidirectional Encoder Representations from Transformers) [3] is a Transformer-approach, pre-trained on large corpora and open-sourced.

As input to the network, we concatenate the sentence and the topic (separated by the [SEP]-token), as follows: *[CLS] The researchers say the yearly flu shot targets protein "heads" that attack the body and make people feel sick. [SEP] Public health demands vaccinations*.
We add a softmax layer to the output of the first token from BERT and fine-tune the entire *bert-base-uncased model* ($\text{BERT}_{base}$) and the *bert-large-uncased model* ($\text{BERT}_{large}$) with an Adam optimizer for three epochs with a batch size of 16, a maximum sequence length of 128 and a learning rate of 2e-5. To train our module, we use the UKP Sentential Argument Mining Corpus [14], containing 400 documents with 25.492 sentences on eight controversial topics (abortion, cloning, death penalty, gun control, marijuana legalisation, minimum wage, nuclear energy and school uniforms), annotated with the same three labels: *Argument-InFavor*/*Argument-Against*/*Neither*.

Given that the purpose of the proposed stance detection module is to be integrated into a disinformation analysis tool, to better evaluate its performances in the targeted context, we have collected and annotated a sample of fake news, and we have annotated them with the stance labels described before. The following section describes the dataset construction.

### 3.1 Test Set Creation

We randomly selected a set of articles identified as containing fake information on a given target topic by the Storyzy fake news detection system (Section 2). In total, we collected 86 articles containing nearly 3000 sentences (after a pre-processing phase

aimed at removing very short and useless sentences) and equitably covering three current controversial topics: White Helmets provide essential services, Public health demands vaccinations and The impact of Covid-19 is risible.[6] For the annotation phase, we followed the same annotation scheme and protocol used to annotate the UKP Corpus [14]. Our expert annotators were three researchers in the area of stance detection and fake news detection, whose goal was to assign to each sentence one of the three stance labels, i.e., *Argument-InFavor*, *Argument-Against* and *Neither*. We would like to recall that our goal is not to annotate only arguments which contain false information about a given target topic, but rather to annotate all arguments (i.e., those which contain false information or not) related to a given topic in the article. Table 1 provides some examples of argumentative sentences and the assigned stance labels. Inter-annotator agreement (IAA) among the three annotators calculated on 100 arguments is 0.71 (Fleiss'$\kappa$).

| **Target topic** | **Argument** | **Stance** |
|---|---|---|
| *Public health demands vaccinations* | But the reality is that vaccines are loaded with chemicals that destroy immunity, damage the cellular system, and in some cases even result in sterilization. | Argument-Against |
| *Public health demands vaccinations* | According to the GreenMedinfo website, the push for the flu vaccine is primarily economic and political rather than based on solid medical evidence. | Argument-Against |
| *White Helmets provide essential services* | In Syria we are seeing the unprecedented use of children by the white helmets as propaganda tools to promote a humanitarian war to kill more children. | Argument-Against |
| *The impact of Covid-19 is risible* | Covid-19 is not serious because the pandemic can be dramatically slowed, or stopped, with the immediate widespread use of high doses of vitamin C. | Argument-InFavor |
| *Public health demands vaccinations* | Vaccines are one of the biggest public health victories in human history. | Argument-InFavor |
| *White Helmets provide essential services* | They [White Helmets] are working very hard in a very dangerous situation, doing something few others could do... | Argument-InFavor |
| *Public health demands vaccinations* | Georgia State is now looking to move their tests of the nanoparticle vaccine on to ferrets, who have a similar respiratory system to humans. | Neither |
| *The impact of Covid-19 is risible* | PM Sanchez has announced the government will hold meetings via video conference after fellow minister Irene Montero tested positive for the virus. | Neither |

**Table 1.** Examples of argumentative sentences found in fake news.

[6] The dataset of fake news annotated with stance labels is available at `https://github.com/jeris90/annotationFN`.

Table 2 provides statistics on the size and the class distribution of our data set. A first observation is that the proportion of topic-related arguments in our dataset is generally lower than the proportion observed in the UKP corpus. Indeed, our corpus (resp. the UKP corpus) contains 82.6% (resp. 56.3%) of sentences with the label "Neither", 13.1% (resp. 24.3%) for the label "Argument-Against", and 4.3% (resp. 19.4%) for the label "Argument-InFavor". There are several explanations for this difference. First of all, while the data in the UKP corpus generally comes from sources that focus on the debate on a given topic (e.g., `https://www.procon.org`), the articles containing fake news are rarely completely focused on a single topic, thus increasing the number of sentences with the label "Neither". A second observation is the very low number of arguments in favor of the target topic. Overall, in fake news, arguments in favor of the target topic are often arguments from "certified" articles or web sites which are attacked (and therefore discredited) with the goal of destabilising the society.

| topic | articles | sentences | Argument-InFavor | Argument-Against | Neither |
|---|---|---|---|---|---|
| white helmets | 25 | 998 | 20 | 110 | 868 |
| vaccination | 31 | 942 | 84 | 152 | 705 |
| covid-19 | 30 | 1008 | 24 | 123 | 861 |
| total | 86 | 2947 | 128 | 385 | 2434 |

**Table 2.** Topics, corpus size and label distribution.

### 3.2 Results and Error Analysis

In addition to the two BERT models (i.e., $BERT_{base}$ and $BERT_{large}$), we also compared our results with a majority baseline (i.e., the prediction is the most common label in the training dataset). Table 3 reports on the results returned by these three stance detection models on our fake news test set. The fine-tuned $BERT_{base}$ performs 0.44 and 0.04 better in $F_1$ score than the majority baseline and $BERT_{large}$, respectively. Best results are obtained for both classes (i.e., *Argument-InFavor* and *Argument-Against*) with $BERT_{base}$.

**Error Analysis.** Table 4 provides the confusion matrix of $BERT_{base}$. The reasons for some misclassifications are firstly due to the lack of word knowledge that can bias the perception of the polarity of a given argument (argument in favor or against). This is the case, for example, for elements with a positive connotation such as prices/rewards (e.g., the sentence *"The Nobel Peace Prize must go to the White Helmets."* is labelled *Neither* while this is an argument in favor of White Helmets), or for elements with a negative connotation such as terrorists (e.g., the sentence *"The well-known organization White Helmets once again stepped up its activities in Syria, enlisting, by tradition, the support of one of the largest terrorist groups."* is labelled *Neither* while it is an argument denouncing the relations of White Helmets with terrorists whilst they are supposed to be a humanitarian organisation). A second problem is related to the ambiguity

| Model | $\mathbf{P_{arg_+}}$ | $\mathbf{P_{arg_-}}$ | $\mathbf{P_{narg}}$ | $\mathbf{R_{arg_+}}$ | $\mathbf{R_{arg_-}}$ | $\mathbf{R_{narg}}$ | $\mathbf{F_1}$ | **Accuracy** |
|---|---|---|---|---|---|---|---|---|
| majority baseline | 0 | 0 | 0.83 | 0 | 0 | 1 | 0.30 | 0.83 |
| $\text{BERT}_{base}$ | 0.50 | 0.63 | 0.97 | 0.77 | 0.77 | 0.91 | 0.74 | 0.89 |
| $\text{BERT}_{large}$ | 0.55 | 0.62 | 0.93 | 0.63 | 0.55 | 0.93 | 0.70 | 0.87 |

**Table 3.** Results of each model on our Fake News Dataset, where **P** is for precision, **R** for recall, $\mathbf{arg_+}$ for the label *Argument-InFavor*, $\mathbf{arg_-}$ for the label *Argument-Against* and **narg** for the label *Neither*.

and subjectivity of the arguments. For example, the following argument *"Therefore it is logical to assume that the White Helmets are aiding the U.S. government to achieve these aims and have been handsomely bankrolled for their efforts."* is erroneously classified as *Argument-InFavor* by the system, while it is labelled as *Argument-Against* in our gold standard because the author, with this sentence, denounces the connection of White Helmets with the U.S. Government whilst they are supposed to be an independent organisation.

| | | **Predicted label** | | |
|---|---|---|---|---|
| | | Argument-Against | Argument-InFavor | Neither |
| **Gold label** | Argument-Against | 295 | 41 | 49 |
| | Argument-InFavor | 18 | 99 | 11 |
| | Neither | 161 | 58 | 2215 |

**Table 4.** Confusion matrix on the test set for the $\text{BERT}_{base}$ model.

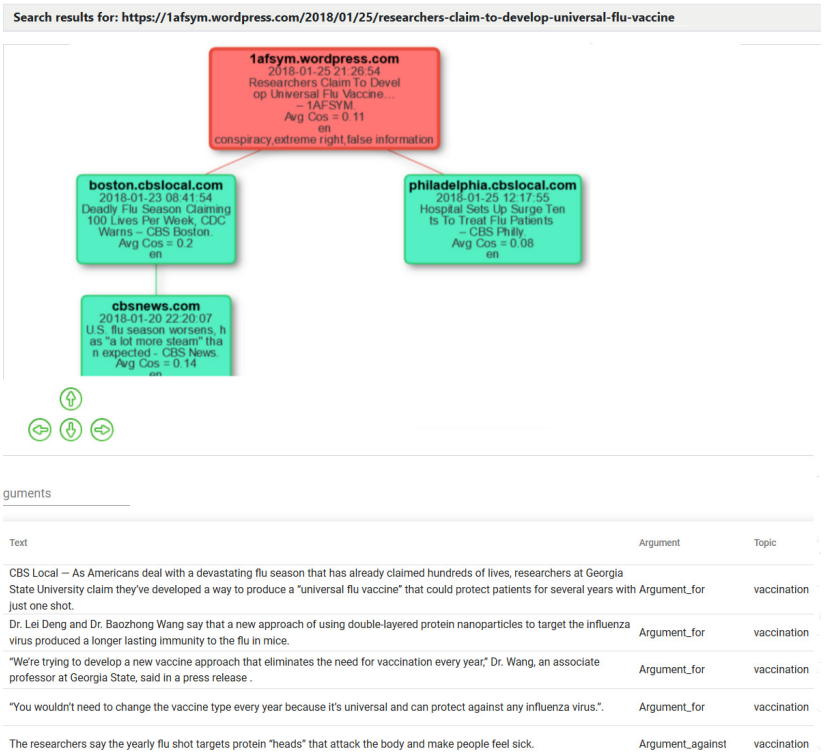
## 4 Integrating Stance Detection in the Disinformation Analysis Tool

Storyzy Disinformation Analysis Tool allows users to investigate disinformation diffusion according to several criteria:

- Topic models are built to identify relevant key-words for a given set of documents,
- The authors of these documents are extracted and a network is built to represent their contributions to different information sources,
- A chronological graph is built to identify the source of the disinformation and its current stage of diffusion,
- Incoming and outcoming links are analyzed and categorized according to their reliability, allowing the system to give scores to information sources according to the nature of these links.

These elements provide insightful information to aid the analyst in her task. For instance, it is possible to identify some recurrent networks of specific information sources allowing the analyst to better understand how some information emerged and began to spread across the internet. This then makes it possible to automatically generate (when possible) the propagation path of an information, a task that has been done manually until now.[7]

**Fig. 1.** Screenshot of the Storyzy Disinformation Analysis Tool showing, for the text located in the URL at the top of the figure, its "chronological" graph and a list of arguments in favor and against the target topic of "Public health demands vaccinations" returned by the stance detection module.




The interaction between the Disinformation Analysis Tool and the stance detection module is done via an API. The latter requests the following three elements:

1. the plain text of the article;
2. the target topic for which we wish to extract the arguments (this information is available thanks to the set of target topics associated with each article by the DAT);

[7] An example may be found on the website `https://www.newsguardtech.com/covid-19-myths/` with some of the most popular fake news about Covid-19.

3. and the entire URL pointing to the website where the text was extracted (optional).

After checking the received information, the text is split into sentences, converted to the BERT format and the stance detection module is run to assign a label ("Neither", "Argument-InFavor" or "Argument-Against") to each of these sentences. The data is then returned to the DAT in JSON format where each item is associated with a sentence. More precisely, each item contains:

- the plain text of the sentence;
- its label;
- the target topic used by the model;
- and its source (if available).

These elements are then directly integrated into the Disinformation Analysis Tool on the page containing the information about this article in the form of a table containing the sentences that have been labelled as *Argument-InFavor* or *Argument-Against* by our module. We decided not to display the sentences being labelled as *Neither*, as they are not of help for the analysis. Figure 1 shows a screenshot of the Storyzy Disinformation Analysis Tool empowered with the stance detection module.

The URL of this article is at the top of the figure. The chronological graph of this article is schematised, showing the links (i.e., citations) of this article to other articles and vice versa. The nodes represent the different sources involved. The red rectangles identify the unreliable sources, the green ones the reliable sources and the grey ones (when available) the neutral sources. An arrow between two nodes means that the articles in the top node quote the articles represented in the bottom node. In other words, the articles at the top of the graph are the most recent ones, while those at the bottom are the oldest. This allows us to easily follow the spread of information related to this article and its chronological evolution.

Finally, at the bottom of the page, we find the list of the sentences identified as arguments by our module for the same article. In this case, our module has identified five arguments related to the target topic "Public health demands vaccinations". Of these arguments, seven were labelled "Argument-InFavor" (e.g.,*"Dr. Lei Deng and Dr. Baozhong Wang say that a new approach of using double-layered protein nanoparticles to target the influenza virus produced a longer lasting immunity to the flu in mice."*) and one was labelled "Argument-Against" (i.e., *"The researchers say the yearly flu shot targets protein heads that attack the body and make people feel sick."*).

The list of arguments associated with each article returned by our model is ready to be used by fact-checkers and journalists who are interested in articles containing fake information on a specific target topic. The newly proposed tool offers indeed the possibility to, for instance, find "popular" articles, thanks to the chronological graph. This graph supports discovering the article at the origin of a fake news, but also providing a (total or partial) overview of the fake news propagation and thus knowing which article(s) played an important role in the propagation of this information. Having a list of arguments in favor or against the target topic of these articles allows the user to quickly highlight and analyse the kind of argument used by the author of the article to potentially convince the reader.

## 5 Conclusions

In this paper, we present an approach to sift the arguments in fake news to boost a disinformation analysis tool, with the final goal of supporting fact-checkers and data analysts in detecting fake news online and in identifying the precise fake information spread through the article. As a side contribution, we have introduced a new annotated resource for stance detection from articles containing fake news.

Several improvements regarding the practical use of this list of arguments are considered for future work. First, establishing statistics related to arguments in fake news. Indeed, the number of fake news can potentially vary from one target topic to another and can be extremely high for highly controversial topics such as adoption, vaccination, etc. The Disinformation Analysis Tool being able to provide the list of articles identified as coming from an unsafe source for a given topic, it would be interesting to make an overall assessment of the arguments extracted from these articles. This can, for example, take the form of a list of the X most commonly used arguments in fake news for a given topic. In order to create such list, it is necessary to be able to compute clusters of arguments according to different criteria, the main one being the degree of similarity between these arguments.

Second, we are currently investigating how the arguments and their stance can improve the automatic detection of articles containing false information. Indeed, we are interested to see whether argument-related criteria such as the number of arguments or the stance of the identified arguments can be used to "facilitate" the automatic detection of fake news. To go even further, it would be interesting to have a graphical representation of the arguments and their interactions within the same article. This would, for example, reveal potential inconsistencies in the article and thus reveal misleading fallacious argumentation.

Third, establishing a counter-argument process to convince those who unintentionally share this false information that some of the arguments put forward are fallacious. Counter-argumentation [8] is a process aiming to put forward counter-arguments in order to provide evidences against an existing argument. In the case of fake news, in order to convince a person that the (fake) information is true, the author of the fake news will use different methods of persuasion via arguments. Thus, identifying these arguments and attacking them by using carefully constructed arguments from safe sources is a way to fight this phenomenon and its spread on social networks. This means that it would be necessary to classify arguments in order to understand which ones are worth for counter-argumenting. Some criteria such as verifiability (verifiable vs. unverifiable) [10] or factuality (facts vs. opinions) [16] can be used for this purpose. Other arguments from sources deemed reliable by the tool can also be used as a counter-argument.

Finally, we plan to address a user study to evaluate the effectiveness of the user interface presenting the arguments and their stance in fake news articles.

## 6 Acknowledgements

This work benefited from the support of the project DGA RAPID CONFIRMA.

## References


1. Elena Cabrio and Serena Villata. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, pages 5427–5433, 2018.
2. Claudette Cayrol and Marie-Christine Lagasquie-Schiex. On the acceptability of arguments in bipolar argumentation frameworks. In *Proceedings of the 8th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU'05*, pages 378–389, 2005.
3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *http://arxiv.org/abs/1810.04805*, 2018.
4. Neema Kotonya and Francesca Toni. Gradual argumentation evaluation for stance aggregation in automated fake news detection. In *Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL 2019, Florence, Italy, August 1, 2019*, pages 156–166, 2019.
5. Dilek Küçük and Fazli Can. Stance detection: A survey. *ACM Comput. Surv.*, 53(1), February 2020.
6. John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818, 2019.
7. Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.*, 16(2):10:1–10:25, 2016.
8. Hugo Mercier and Dan Sperber. Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74, 2011.
9. Ray Oshikawa, Jing Qian, and William Yang Wang. A survey on natural language processing for fake news detection. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6086–6093. European Language Resources Association, 2020.
10. Joonsuk Park and Claire Cardie. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argument Mining, hosted by the 52nd Annual Meeting of the Association for Computational Linguistics, ArgMiningACL'14*, pages 29–38, 2014.
11. Antonio Rago, Francesca Toni, Marco Aurisicchio, and Pietro Baroni. Discontinuity-free decision support with quantitative argumentation debates. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning, KR'16*, pages 63–73, 2016.
12. Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 567–578, 2019.
13. Ricky J. Sethi. Spotting fake news: A social argumentation framework for scrutinizing alternative facts. In *Proceedings of the 2017 IEEE International Conference on Web Services, ICWS'17*, pages 866–869, 2017.
14. Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*, pages 3664–3674, 2018.

15. Rubén Tolosana, Rubén Vera-Rodríguez, Julian Fiérrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *CoRR*, abs/2001.00179, 2020.
16. Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2003*, 2003.
17. Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM'19*, pages 836–837, 2019.