Identification of Fake News by Contradiction Detection in Texts

Identificación de Noticias Falsas mediante Detección de Contradicciones en Textos

Robiert Sepúlveda-Torres
Departamento de Lenguajes y Sistemas Informáticos,
Universidad de Alicante, E-03080 Alicante, Spain
rsepulveda@dlsi.ua.es

Abstract: The dissemination of fake news through digital media has increased significantly in recent years. The volume of generation of this kind of news is so high that it is impossible to verify them manually, being necessary to use technologies that allow automating the verification process. This work is focus on creating models and technologies that allow supporting the fake news detection process. The main advances in the area are showed, as well as planning to carry out the research. Finally, the results obtained so far in the research are explained.
Keywords: Natural Language Processing, Fake news, Fact-Checking, Stance detection, Contradiction detection

Resumen: La diseminación de noticias falsas a través de medios digitales ha aumentado significativamente en los últimos años. El volumen de generación de estas noticas es tan alto que es imposible su verificación manual siendo necesario usar tecnologías que permitan automatizar el proceso de verificación. Este trabajo se enmarca en crear modelos y tecnologías que permitan apoyar el proceso de detección de noticas falsas. Se plantean los principales avances en el área, así como una planificación para llevar a cabo la investigación. Por último, se explican los resultados obtenidos hasta el momento en la investigación.
Palabras clave: Procesamiento de lenguaje natural, Noticias falsas, Chequeo automático de hechos, Detección de posturas, Detección de contradicciones

1 *Introduction and motivation*

Low cost and rapid access to digital media and social networks have led to increased consumption of digital content on these platforms (Conroy, Rubin, and Chen, 2015; Rubin and Lukoianova, 2015). These platforms, mainly blogs and social networks, are not validated and are therefore conducive to the proliferation of fake news (Shu et al., 2017).

Fake news has existed for a long time (Allcott and Gentzkow, 2017), but the term "fake news" is relatively new, and it was defined by The New York Times as a "made up story with the intention to deceive, often with monetary gain as a motive" (Tavernisen, 2019). This phenomenon has experienced a significant boom since the 2016 US election (Bovet and Makse, 2019) and the Brexit referendum 2016 (Bastos and Mercea, 2019). In addition, according to (Vosoughi, Roy, and Aral, 2018) fake news is 70% more likely to be shared than real news so the use of these platforms has become a double-edged sword (Shu et al., 2017).

The amount of fake news generated and distributed by digital media every day is very high, hence the manual evaluation of its veracity is practically impossible in a reasonable time frame (Tsipursky, Votta, and Roose, 2018). In recent years, Artificial Intelligence techniques have been used to support the fake news detection process (Andreas Hanselowski and Caspelherr, 2017; Thorne et al., 2018; Rubin and Lukoianova, 2015; Conroy, Rubin, and Chen, 2015).

In (Saquete et al., 2020) a systematic review of the fake news phenomenon was conducted. They identify the main tasks that can intervene in the fake news detection process, such as: **Deception detection**; **Stance detection, controversy and**

polarization; **Automated fact-checking**; **Clickbait detection** and **Credibility**. An open problem in this research area is the challenge of integrating the independent tasks in the fake news detection in a complex process.

Another fundamental task in the fake news detection process is identifying contradictions in texts, transversely, in all the other tasks involved in fake news detection. Contradiction detection is a complex task in Natural Language Processing due to the variety of ways that it manifests itself between texts (Lingam et al., 2018). The main forms of contradiction are defined in (de Marneffe, Rafferty, and Manning, 2008) as: **Antonym**; **Negation**; **Numeric**; **Factive**; **Factive Structure**; **Structure** and **Lexical**.

The research presented in this paper is focused on fake news detection by using computational models that allow the identification of cues of falsehood in the evaluated news. Contradiction detection between texts is sharpened, which is a fundamental way of supporting the fake news detection task. Finally, the aim is to integrate the main tasks that interfere with the fake news detection process with a contradiction detection module that can be used cross-sectionally throughout the fake news detection process.

2 Background and Related Work

In this section, an in-depth review of existing fake news detection strategies is conducted. The review is based on the main tasks within fake news detection.

The fake news detection is usually done by obtaining the linguistic features (Gravanis et al., 2019; Chua and Banerjee, 2016) or by evaluating the context of the news (Shu et al., 2018). With these features, systems based on machine learning are created to carry out the detection.

There are some organizations that are engaged in fact-checking such as Snopes[1], FactCheck[2] and Newtral[3]. They usually have a group of fact-checkers who verify the facts manually. There are attempts to automate the task of fact-checking and stance detection with some workshops and challenges such as Fact Extraction and Verification (FEVER)[4] and Fake News Challenge (FNC-1)[5] that try to deepen in approaches based on Natural Language Processing, Machine Learning and Deep Learning. In FEVER (Thorne et al., 2018), a corpus is developed for automatic fact-checking. In the past editions, about 30 systems have been developed, some of them obtaining very high scores. The winning team (Stammbach and Neumann, 2019) in this challenge proposed a system based on BERT (Devlin et al., 2018), which recovers similar sentences in two stages obtaining more precise evidence for the final classification.

The FNC-1 is a stance detection challenge that consists of estimating the relative perspective (or stance) of two pieces of text in relation to a topic, complaint or issue (Riedel et al., 2017). The three best performing systems in this competition were Talos (Baird, Sibley, and Pan, 2017), Athene system (Andreas Hanselowski and Caspelherr, 2017) and UCLMR (Riedel et al., 2017), respectively. The winning team (Talos) (Baird, Sibley, and Pan, 2017) used an ensemble model based on a 50/50 weighted average between gradient-boosted decision trees and deep convolutional neural networks (CNN) on the headline and body text, represented at the word level using Google News pre-trained vectors.

When a news item is misleading, it introduces contradictory information to a true news item and therefore the detection of contradictions is a fundamental task when you want to identify with fake news. Contradiction detection is the task of identifying pairs of natural language statements, conveying information about events or actions that cannot simultaneously hold (Dragos, 2017).

The most common state-of-the-art approaches for contradiction detection in text is to use linguistic features extracted from text to build a classifier and train from annotated examples (Lingam et al., 2018; Lendvai and Reichel, 2016). In (de Marneffe, Rafferty, and Manning, 2008), the types of contradictions that can be found between texts are detailed and an approach is proposed to detect a subset of these.

(Lingam et al., 2018) proposed an approach for detecting three different types of contradiction: negation, antonyms, and numeric mismatch. This approach adopts a Re-

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

[2] https://www.factcheck.org/

[3] https://www.newtral.es/

[4] https://fever.ai/

[5] http://www.fakenewschallenge.org/

current Neural Network (RNN) using Long short-term memory (LSTM) and Global Vectors for Word Representation (GloVe) and includes four linguistic features extracted from text (jaccard coefficient, negation, is antonym, overlap coefficient). Similarly, (Lendvai and Reichel, 2016) utilized simple text similarity metrics (cosine similarity, f1 score and local alignment) that, as baseline, obtain a good result for contradiction classification.

The state of the art of contradiction detection shows that although the problem is well defined, clarity is lacking in the methodologies adopted and, furthermore, the results obtained do not indicate high precision. Furthermore, most of the work is done with their own datasets created to test the proposed models, so the contradiction detection suffers from a lack of a gold standard dataset to compare the results of the proposed models.

3 Main Hypothesis and Objectives

The PhD thesis presented aims to obtain a generic architecture for fake news detection based on the contradiction detection and its integration into the other main tasks involved in fake news detection. The main hypothesis of this research is that it is possible to formalize a model which enables the contradiction detection between texts, and that said model can be integrated into the fake news detection process. The objectives proposed to carry out the research are to:

- Examine thoroughly the open problems within the fake news detection process.
- Identify the main elements that must be taken into account to detect contradictions in texts.
- Propose a generic model for detecting contradictions in texts that can be integrated into the fake news detection process.
- Propose different methodologies based on the generic model.
- Propose a detection architecture that integrates the tasks of fake news detection and validates the relevance of the contradiction detection model.
- Validate the fake news detection architecture.

4 Methodology and the proposed experiments

The methodology adopted to achieve the results in this research is based on a systematic study of the state of the art. An analysis of scientific papers and material related to fake news detection and contradiction detection is conducted to identify the most important open problems in this area of research, for which solutions are proposed. In the first phase, resources, methodologies and tools will be gathered to support the research process in the area. This process must be iterative because techniques usually change in a short time frame.

A step-by-step process is proposed that will guide the research at each stage and can be viuslaized in Figure 1:

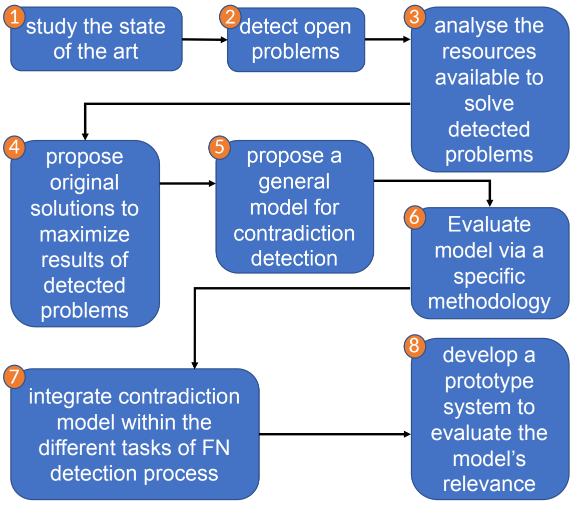
1 study the state of the art
2 detect open problems
3 analyse the resources available to solve detected problems
4 propose original solutions to maximize results of detected problems
5 propose a general model for contradiction detection
6 Evaluate model via a specific methodology
7 integrate contradiction model within the different tasks of FN detection process
8 develop a prototype system to evaluate the model's relevance

Figure 1: Step-by-step process

Where relevant, results will be published in scientific journals and papers will be presented at conferences so as to disseminate the advances made in this ongoing research.

Both automated fact-checking and stance detection are two sub-tasks within fake news detection and currently, work has been done on both as described below. So far the research has been conducted using English language resources, in the future, we intend to include other languages such as Spanish.

4.1 Automated fact-checking

An automated fact-checking model using the FEVER corpus (Thorne et al., 2018) is proposed in (Alonso-Reina et al., 2019). This approach consists of determining the relationship between a claim and evidence extracted from a knowledge base. The FEVER corpus is composed of a set of claims and their respective classifications (Supports, Refutes and Not Enough Info). The proposed model contains three modules:

- **Document retrieval**, the main goal is to obtain relevant pages, using Wikipedia as a knowledge base. This task retrieves those pages containing elements related to the claim under evaluation.
- **Sentences retrieval**, the sentences most similar to the claim are extracted from the documents obtained from the knowledge base in the previous module. A similarity value between the triplets (Subject, Verb, Object) of each sentence is calculated with that of the claim. They are placed in descending order of similarity and a maximum of 5 sentences are used to make the final classification in the last module.
- **Recognizing textual entailment**, in charge of classifying the claim with the most similar sentences. A classifier based on recurrent neural networks is created using the ESIM model (Chen et al., 2017) to obtain an inference between the claim and the sentences. The inputs in this classification module are embedding vectors of the claim and sentences.

This approach works well in the FEVER challenge and its methodology can be generalized even by replacing its models with more powerful ones.

4.2 Stance detection

The stance detection task consists of determining the different positions in discussions among different people, this allows the classifying of the comments into groups and evaluation of their veracity (Saquete et al., 2020). The main works in this task have been developed on a corpus of a significant size called Fake New Challenge (FNC-1). This corpus contains news bodies and their possible headlines to be classified in (Agree, Disagree, Discuss, Unrelated). This corpus is used in the next two research projects.

The main approaches that obtain better results for stance detection are based on neural models but these models can have a negative impact on efficiency when processing long texts (Yoon et al., 2018). In (Hayashi and Yanagimoto, 2018), the first sentence of the text or a specific fragment (Huang et al., 2017) has been used to combat this problem. Another option is to use summary algorithms that could be beneficial in this context.

Our first study compares the impact of some summary algorithms in the performance of stance detection models. Different types of summary approaches, as well as two stance detection methods, based on machine learning (Ferreira and Vlachos, 2016) and deep learning (Chen et al., 2017), are tested on two state-of-the-art datasets, Emergent (Ferreira and Vlachos, 2016) and FNC-1. The Emergent dataset is similar to FNC-1 dataset. For this dataset, the classes of the dataset are (For, Against, Observing). The results obtained corroborate that basically when the text is very long, the use of summaries is a valid option that usually improves the results obtained by the machine learning models and deep learning models compared to the full text. These results are in line with the summarizer technique used, and generally, extractive summaries work better. In our case, the best results were achieved with the PLM Summarizer (Vicente, Barros, and Lloret, 2018).

Our second study in this task is based on creating a stance detection architecture that includes the use of PLM Summarizer with external features that allow for improved classification results. The corpus of FNC-1 is used. The architecture used the divide and conquer strategy to first classify elements into Related and Unrelated and then Related are classified into Agree, Disagree and Discuss.

The architecture is composed of two stages (Relatedness Stage and Stance Stage). In the relatedness stage, we determine whether a headline of an article is similar to the body text. This stage contains a summary module that allows the summarization of the body text into a high-quality summary. Another module that contains the relatedness stage is the relatedness feature extraction that allows for the calculation of some distance and similarity metrics. Finally, the summary and the headline plus the extracted features are used to carry out the classification. The second stage of the architecture —stance stage— receives the summary and the headline of the article and performs the classification. Each stage contains a classification module that allows each classification to be carried out. It is important to mention that this architecture is divided into stages and each stage into modules, allowing the architecture to be scalable. The modules can be replaced by others with better results and even new modules may be added.

This work improves the state of the art on this corpus. New learning strategies and discourse aware techniques will help to combat online fake news, a societal problem that requires concerted action.

5 Discuss

The research carried out thus far has reached completion on Step 5 (see Figure 1), whereby original solutions have been proposed to maximize results of the detected problems in the fake news detection process.

In the near future, progress is expected to be made on the pending steps that will allow the proposed objectives to be met, the hypothesis to be validated, and new lines of research to be opened in this area. It is necessary to expand research in other languages —for example, in Spanish— to identify whether corpora in other languages are needed to create methodologies for fake news detection, or whether cross-lingual approaches can solve these problems adequately.

The ultimate goal is to generate a powerful contradiction detection model that will be useful in the process of detecting fake news. This model will have the capacity to be fully integrated into more complex fake news detection processes.

Acknowledgements

This research work has been partially funded by the University of Alicante (Spain), Generalitat Valenciana and the Spanish Government through the projects Tecnologías del Lenguaje Humano para una Sociedad Inclusiva Igualitaria y Accesible (PROMETEU/2018/089), Modelado del Comportamiento de Entidades Digitales mediante Tecnologías del Lenguaje Humano (RTI2018-094653-B-C22) and Integer: Intelligent Text Generation, Generación Inteligente de Textos (RTI2018-094649-B-I00).

References

Allcott, H. and M. Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Alonso-Reina, A., R. Sepúlveda-Torres, E. Saquete, and M. Palomar. 2019. Team GPLSI. approach for automated fact checking. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 110–114, Hong Kong, China. Association for Computational Linguistics.

Andreas Hanselowski, Avinesh PVS, B. S. and F. Caspelherr. 2017. Description of the system developed by team athene in the FNC-1. https://github.com/hanselowski/athene_system , last accessed on 29/05/20 .

Baird, S., D. Sibley, and Y. Pan. 2017. Talos targets disinformation with fake news challenge victory. https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html , last accessed on 29/05/20 .

Bastos, M. T. and D. Mercea. 2019. The brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, 37(1):38–54.

Bovet, A. and H. A. Makse. 2019. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):1–14.

Chen, Q., X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen. 2017. Enhanced LSTM for Natural Language Inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chua, A. Y. and S. Banerjee. 2016. Linguistic predictors of rumor veracity on the Internet. *Lecture Notes in Engineering and Computer Science*, 1:387–391.

Conroy, N. J., V. L. Rubin, and Y. Chen. 2015. Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, page 82. American Society for Information Science.

de Marneffe, M.-C., A. N. Rafferty, and C. D. Manning. 2008. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio, June. Association for Computational Linguistics.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (Mlm).

Dragos, V. 2017. Detection of contradictions by relation matching and uncertainty assessment. In *Procedia Computer Science*, volume 112, pages 71–80. Elsevier B.V.

Ferreira, W. and A. Vlachos. 2016. Emergent: a novel data-set for stance classification. pages 1163–1168.

Gravanis, G., A. Vakali, K. Diamantaras, and P. Karadais. 2019. Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, 128:201–213.

Hayashi, Y. and H. Yanagimoto. 2018. Headline generation with recurrent neural network. In *New Trends in E-service and Smart Computing*. Springer, pages 81–96.

Huang, Z., Z. Ye, S. Li, and R. Pan. 2017. Length adaptive recurrent model for text classification. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1019–1027.

Lendvai, P. and U. Reichel. 2016. Contradiction detection for rumorous claims. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)*, pages 31–40, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Lingam, V., S. Bhuria, M. Nair, D. Gurpreetsingh, A. Goyal, and A. Sureka. 2018. Deep learning for conflicting statements detection in text.

Riedel, B., I. Augenstein, G. P. Spithourakis, and S. Riedel. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *Computing Research Repository, CoRR*, abs/1707.03264.

Rubin, V. L. and T. Lukoianova. 2015. Truth and deception at the rhetorical structure level. *Journal of the Association for Information Science and Technology*, 66(5):905–917.

Saquete, E., D. Tomás, P. Moreda, P. Martínez-Barco, and M. Palomar. 2020. Fighting post-truth using natural language processing: A review and open challenges. *Expert Systems with Applications*, 141:112943.

Shu, K., D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *CoRR*, abs/1809.01286.

Shu, K., A. Sliva, S. Wang, J. Tang, and H. Liu. 2017. Fake News Detection on Social Media. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Stammbach, D. and G. Neumann. 2019. Team DOMLIN: Exploiting evidence enhancement for the FEVER shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 105–109, Hong Kong, China, November. Association for Computational Linguistics.

Tavernisen, S. 2019. As fake news spreads lies, more readers shrug at the truth. *New York Times*.

Thorne, J., A. Vlachos, C. Christodoulopoulos, and A. Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, pages 809–819.

Tsipursky, G., F. Votta, and K. M. Roose. 2018. Fighting Fake News and Post-Truth Politics with Behavioral Science: The Pro-Truth Pledge. *Behavior and Social Issues*, 27(1):47–70.

Vicente, M., C. Barros, and E. Lloret. 2018. Statistical language modelling for automatic story generation. *Journal of Intelligent & Fuzzy Systems*, 34(5):3069–3079.

Vosoughi, S., D. Roy, and S. Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Yoon, S., K. Park, J. Shin, H. Lim, S. Won, M. Cha, and K. Jung. 2018. Detecting Incongruity Between News Headline and Body Text via a Deep Hierarchical Encoder. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:791–800.