

# Veracity vs. Reliability: Changing the Approach of Our Annotation Guideline

Alba Bonet-Jover

*Department of Software and Computing Systems, University of Alicante, Spain*

## Abstract

This paper presents the evolution of an annotation guideline designed for the disinformation detection task, an essential step of my doctoral thesis. The annotation proposal aims to label all the structural and content elements of a news item, as well as to classify them as Reliable or Unreliable. The initial objective was to annotate those elements into Fake or True, but for that classification, world knowledge is needed. Our current goal is to annotate news on the basis of a purely textual, semantic and linguistic analysis, without using external knowledge and, for that reason, the annotation was redirected towards a reliability rating, rather than a veracity classification. This article justifies the change of perspective at this stage of the thesis, explains the difference between veracity and reliability and shows the concrete changes that have been adopted in our annotation proposal with this new approach.

## Keywords

Natural Language Processing, Human Language Technologies, Disinformation Detection, Corpus Annotation, Annotation Guideline

## 1. Justification of the research

Disinformation has become a global social problem. As defined by [1], disinformation is information that is intentionally misleading and that is likely to cause people to hold false beliefs. Fake news is one of the most widespread and well-known disinformation phenomena. The main hypothesis of our research is that fake news usually mixes both true and false information, so if all the parts of a news item are annotated with a single veracity value, it is possible to obtain the global veracity value of the news item. This annotation may allow to train a system that automatically detects the deception of these parts. A system capable of detecting disinformation may help users to be aware of those elements that make a news item not credible. It is not intended to be a categorical tool, but a support tool, so the aim is not that the system decides whether something is completely true or false, but to guide users to build their own opinion. The human has the final decision.

We designed an annotation guideline focused on digital news annotation and based on two well-known journalistic techniques: the Inverted Pyramid and the 5W1H. In the initial stage of the thesis, the labels were created taking into account the main purpose of our research: to

---


*Doctoral Symposium on Natural Language Processing from the PLN.net network 2022 (RED2018-102418-T), 21-23 September 2022, A Coruña, Spain.*

✉ alba.bonet@dlsi.ua.es (A. Bonet-Jover)

🆔 0000-0002-7172-0094 (A. Bonet-Jover)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

assign a veracity value to each structural and content element of the news item in order to detect falsehood patterns.

However, the objective and methodology of the research have been readjusted over time and, instead of assigning a veracity value, the thesis is focused now on defining a reliability criterion based on linguistic, semantic and textual analysis in order to provide cues and warn readers of the presence of questionable information. This article describes the evolution of our annotation proposal, as well as the changes made up to the current stage of the thesis.

The paper is structured as follows: Section 2 presents an overview of relevant scientific literature concerning annotated corpora; Section 3 briefly justifies the change of approach, explains the difference between the reliability and veracity concepts and presents the main objectives of our research; Section 4 details all the changes made to our annotation in each stage of our research; Section 5 presents some interesting elements for discussion and finally Section 6 presents the conclusions of this research and future work.

## 2. Background and related projects

Before introducing our proposal, this section aims to briefly contextualise our research with the state of the art and thus to justify our contribution. Our research arises from two issues: the disinformation global problem, on the one hand, and the scarcity of annotated corpora to train disinformation detection models using NLP, on the other. To face both challenges, we designed an annotation guideline which enables the detection of disinformation patterns through language. According to the literature consulted, most corpora annotated for the disinformation task (either for fake news detection, fact-checking or stance detection, among others) present a binary classification (true/false) based on a veracity criterion.

In addition to the veracity rating, what most datasets have in common is the overall classification of news, assigning them an overall value of true or false. Some interesting corpora that use this veracity rating are those of [2], [3], [4] or [5]. However, as stated by [6], categorizing all the news into two classes (fake or real) is difficult because there are cases where the news is partially real and partially fake. The addition of truthfulness degrees is usually a solution to address this problem, as in the case of the LIAR [7] or the EMERGENT [8] datasets. Even if we annotate with a binary classification of reliability (Reliable/Unreliable), the novelty of our proposal is the individual annotation of each structural and content element separately which allows to differentiate all those elements that mix reliable and unreliable information and to later justify the global value assigned.

Concerning the reliability classification, recently adopted, instead of assigning a veracity value (Fake/True), we focus on the reliability criterion, explained in Section 3. [9] use this classification but they do not apply it to each relevant element of the news item. To the authors' knowledge, there are few studies applying this novel annotation rating and such research does not use such a deep level of annotation. Last but not least, it should be emphasised the importance of the deceptive language in text. Some research, such as that of [10], reach the conclusion that there are differences in the language of fake and real news. Our proposal focuses on analysing textual and semantic characteristics such as connotative expressions, ambiguity, personal remarks or lack of evidence, among others, since our hypothesis lies in the fact that

there exist deceptive signs that influence subjectivity and credibility. The novelty with respect to other research is the analysis of those linguistic characteristics by means of two journalistic techniques which are usually used to communicate a story in a clear, objective and accurate way (the Inverted Pyramid and the 5W1H), while identifying the reliability/unreliability of each part.

### 3. Description and objectives

Our goal is to define linguistic and textual features that can be learned by the system and automatically detected. Therefore, our research is not focused on establishing a criterion of absolute truth by classifying a news item as true or fake, but rather to determine if there is enough evidence in the way the news item is written to consider it credible or not. The paper aims to describe this change of approach and the new classification adopted based on the reliability criterion, which is more novel and precise. It is more feasible to define whether something is credible or not than to categorise whether something is false or true, in which case other factors, such as reader, context and specially world knowledge, influence the decision. Through only linguistic and textual analysis it is not possible to carry out that verification classification, but it is possible to draw conclusions about whether a news item is reliable or not, regardless of whether its content is true or false.

Even if there exists a slightly difference between the concepts veracity and reliability, the essence of this research depends on that divergence. If we compare both entries in the Cambridge dictionary<sup>1</sup>, the term veracity is defined as “the quality of being true, honest, or accurate”, while the term reliability is described as “the quality of being able to be trusted or believed because of working or behaving well”. The difference lies in the verbs “being” and “being able to”. In the same vein, the Britannica Dictionary<sup>2</sup> makes the difference on that probability of being true in the case of the term reliability: “able to be believed, likely to be true or correct” against the certainty of the term veracity: “the quality of being truthful or honest”. Since our research focuses on providing evidence of disinformation and not on giving an absolute classification of truthfulness, we changed our approach and adapted our annotation guideline with the reliability classification.

With this new approach, the research aims to fulfil the three following objectives:

1. **Creation of a text-only annotation:** based on our hypothesis that disinformation can be reflected through language, we propose a purely linguistic and semantic annotation (without using world knowledge) that enables the detection of textual elements that can make a news item unreliable.
2. **Adoption of a reliability criterion:** the change to a reliability-based classification approach may provide readers with disinformation patterns and cues allowing them to evaluate the credibility of a news item before checking the content in official sources. This prior stage allows to quickly get an idea of the news content, the way it is written and its emotional component, generating doubt about its reliability before sharing or believing it.

---

<sup>1</sup><https://dictionary.cambridge.org/es/>

<sup>2</sup><https://www.britannica.com/dictionary>

3. **Creation of a quality dataset:** annotated resources to train in the disinformation task are scarce in NLP, specially in the Spanish language. For that reason, in addition to propose an annotation system, we aim to build a quality dataset specially designed for the disinformation task, with examples linguistically annotated, in order to train our system and make progress in the automatic disinformation detection.

## 4. Methodology

At the beginning of this thesis, the fake news phenomenon was at its peak: news was easily spread and hardly filtered or verified. However, over time, control of disinformation has become increasingly stricter by means of fact-checking agencies, verification sections proposed by digital newspapers and even training courses to raise public awareness of the importance of educating on disinformation. The evolution of this phenomenon has affected our research, changing our objectives and methodology. In this section we justify all the changes and all the different phases undergone by our annotation proposal from the initial hypothesis of the thesis up to now.

### 4.1. First version: FNDeepML

For the first version of our annotation scheme, true news, fake news and fact-checks were manually compiled from digital newspapers. The structure and content elements of each news item were manually annotated in txt format following two well-known journalistic techniques: the Inverted Pyramid (TITLE, SUBTITLE, LEAD, BODY, CONCLUSION) and the 5W1H (WHAT, WHO, WHERE, WHEN, WHY, HOW). Regarding the Inverted Pyramid, the BODY part was not annotated with the 5W1H since it slowed down the annotation, so we decided to leave it unannotated until the annotation was proven in the rest of the parts. All the elements were annotated as True or Fake according to the information provided in the fact-check, or as Unknown if the fact-check did not contrast the data. The Unknown label did not mean that the information was True or Fake, but simply that it cannot be verified. Based on this approach, an annotation guideline named FNDeepML (Fake News Deep Markup Language) was designed ad hoc and validated in a dataset consisting of 200 news (105 True and 95 Fake) built from scratch [11].

During the annotation process, we realised that the annotation in txt format was quite slow and difficult, as the annotated elements were not easily visible, leading the annotator to make many mistakes. For that reason, we decided to change the annotation format and started to use the application GATE Developer 8.6.1<sup>3</sup>, which was configured according to our annotation guideline. Finally, it is important to highlight that, for this first version, news was only labelled according to the information verified by the fact-check, any verification was performed on other external sources. All the fact-checks used belonged to recognised Spanish fact-checking agencies from the International Fact Checking Network (IFCN)<sup>4</sup>.

---

<sup>3</sup><https://gate.ac.uk/download/>

<sup>4</sup><https://www.poynter.org/ifcn/>

## 4.2. Second version: FNDeepML V.2

For this second stage, the annotation scheme was significantly improved and a new approach began to be adopted. The main changes made are described below:

- Change of interface: we started to use the Brat tool, an intuitive web-based annotation tool that aims to enhance annotator productivity by closely integrating NLP technology into the annotation process[12]. Brat is a more intuitive and comfortable interface that allows to quickly and accurately mark the text and that shows labels by means of colours and symbols, which helps the expert annotator in the task. The change of interface has made annotation easier and faster.
- Annotation of the BODY part: in our first version we did not annotate this part, but the BODY contains a lot of information, since it develops and describes in detail the main event of the story. For that reason, we considered it important to annotate in order not to miss key information. The annotation of this part allowed us to obtain more training data for our corpus.
- Creation of new attributes: specific attributes providing additional information about the labels were created for this second version. We can highlight the following attributes: the **main\_event** used with the WHAT label, the **role** to indicate the function of the WHO label (Subject, Target, Both), the **author\_stance** to indicate the agreement (Agree), disagreement (Disagree) or impartiality (Unknown) of the author towards the QUOTE label, the **title\_stance** to mark the coherence between the TITLE and the BODY (Agree, Disagree, Unrelated) and the attribute **type** used to mark if the element annotated is True, Fake or Unknown.

The veracity value of the parts related to the Inverted Pyramid are no longer annotated, since it is very imprecise to manually categorise an entire structural part into True or Fake when inside information mixes both contrasted and false information. For that reason, we established the attribute **title\_stance** to just mark if the information provided in the TITLE and the rest of the news item was contradictory or not.

## 4.3. Third version: RUN-AS

It was becoming increasingly difficult for us to compile fake news for four main reasons: firstly, fact-checking agencies became more involved in the disinformation verification, constantly creating fact-checks that refute fake news and hoaxes; secondly, precisely because of the increased involvement of fact-checking agencies, fake news was better created and camouflaged, adopting the appearance of true news and making it difficult to distinguish reliable information from unreliable information; thirdly, fake news detected was quickly removed or blocked; fourthly, due to the disappearance of fake news in digital newspapers and because of the easy dissemination on social networks, fake content was increasingly spread in form of posts or WhatsApp chains.

In the face of this difficulty of obtaining fake news for our dataset, we had to adapt our methodology to a new stage without changing the objective originally set. It was impossible to obtain a fact-check for each news item collected (because the topic is often so recent or

with such a low impact that it is not interesting to verify it, or simply because the spreading of fake news is faster than the dissemination of verified information). If we add to this the fact that language without context does not give enough information to make a classification and that official external sources are needed to contrast information, it can be concluded that the classification into True or Fake can only be established through fact-checking. However, the fact-checking task is not the purpose of this thesis, so we had to redirect our annotation and change the approach we were following.

For our third version, the one we are using now, we changed the annotation criterion from a veracity classification into True/Fake/Unknown to a reliability rating into Reliable/Unreliable. In addition, the name of our annotation scheme was changed into RUN-AS (Reliable and Unreliable News Annotation Scheme). The purpose is to annotate news into Reliable or Unreliable only by means of linguistic and textual analysis, without relying on external knowledge. This classification would be a prior step to the fact-checking task that may allow to quickly detect suspicious information patterns before contrasting the content with official sources.

Besides replacing the attribute **type** (True/Fake/Unknown) by the attribute **reliability** (Reliable/Unreliable), we created another annotation level, in addition to the Structure (Pyramid Inverted) and Content (5W1H) levels, named Elements of Interest. This third level enables the annotation of additional linguistic and textual characteristics that could help to distinguish unreliable from reliable news. For this level, we created four new labels that are KEY EXPRESSIONS (phraseology with an emotional component), QUOTE, FIGURE and ORTHOTYPOGRAPHY. Furthermore, we added two new attributes to those described in 4.2: **style** (to mark the objectivity or subjectivity of the TITLE) and **lack\_of\_information** (when evidence is missing). In order to test this new annotation approach, a new dataset was created from scratch comprising 85 Reliable news and 85 Unreliable news in Spanish, collected from Spanish digital newspapers and annotated with the Inverted Pyramid, 5W1H and Elements of Interest levels.

## 5. Specific elements of the research for discussion

In this section, we aim to highlight some of the difficulties we have encountered along the way. Firstly, the complexity of our proposal makes the annotation process slow and complicated. A news item is not an archetype presenting information in a rigid form; rather, the style and language vary greatly depending on the journalist, context, topic, recipient or intention. In addition, the analysis of news requires an in-depth linguistic and semantic study which implies expertise, time and coherence. Due to this higher difficulty when annotating, prior training is needed for annotators to learn this annotation proposal. This training could include tutorials and examples of each annotation element. Despite requiring intensive training, the complexity of our annotation is compensated in return by the accuracy of the examples obtained to train.

Another element to be analysed for future work is the simplification of our annotation. One of the advantages of our proposal is that it enables the analysis of many elements and events at the textual and semantic level. However, this exhaustive annotation, in turn, complicates the task. In order for the annotation to be replicated by other researchers, it is necessary to simplify the proposal and to reduce the number of elements and ideas to be labelled. To that end, it would be interesting to extract the main ideas and annotate them following our scheme.

The methodology and the annotation would remain the same, but the text annotated would be simplified and would include the most interesting information to be annotated.

## 6. Conclusions and future work

This paper describes the evolution and the change of approach of my doctoral thesis which focuses on designing an annotation resource for the disinformation detection task. The annotation proposal is based on a linguistic and semantic analysis, as well as on two well-known journalistic techniques. Our last version of the annotation guideline, named RUN-AS, aims to detect all the essential structural and content elements of a news item, to assign them an individual reliability value and thus to obtain a global reliability classification of the news item.

In contrast to the veracity rating, dependent on world knowledge, the reliability classification adopted is more novel and accurate, since it allows to detect suspicious information and disinformation patterns and thus to draw the reader's attention to these features. The purpose is that those linguistic cues show which elements make the information unreliable and make the users hesitate before sharing or believing a news item without evidence.

The objective of describing the change of methodology throughout the thesis is to show that the annotation guideline presented has been carefully designed, tested and continuously updated and modified until achieving a valid resource for our research.

As future work, we are working on an assisted annotation allowing to increase the efficacy and the speed in building our dataset. This proposal is focused on a semi-automatic approach and may allow not only to assist the expert annotator by increasing the performance, but also to create an automated model based on the reliability criterion.

## Acknowledgments

This research work has been partially funded by the Spanish Government and Fondo Europeo de Desarrollo Regional (FEDER) through the project Modelang: Modeling the behavior of digital entities by Human Language Technologies (RTI2018-094653-B-C22) as well as supported by a grant from the Consellería de Innovación, Universidades, Ciencia y Sociedad Digital (ACIF/2020/177) from the Spanish Government.

## References

- [1] D. Fallis, The varieties of disinformation, *The philosophy of information quality* (2014) 135–161.
- [2] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, *arXiv preprint arXiv:1708.07104* (2017).
- [3] J.-P. Posadas-Durán, H. Gómez-Adorno, G. Sidorov, J. J. M. Escobar, Detection of fake news in a new corpus for the spanish language, *Journal of Intelligent & Fuzzy Systems* 36 (2019) 4869–4876.
- [4] R. M. Silva, R. L. Santos, T. A. Almeida, T. A. Pardo, Towards automatically filtering fake news in portuguese, *Expert Systems with Applications* 146 (2020) 113199.

- [5] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an infodemic: Covid-19 fake news dataset, in: International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation, Springer, 2021, pp. 21–29.
- [6] R. Oshikawa, J. Qian, W. Y. Wang, A survey on natural language processing for fake news detection, arXiv preprint arXiv:1811.00770 (2018).
- [7] W. Y. Wang, " liar, liar pants on fire": A new benchmark dataset for fake news detection, arXiv preprint arXiv:1705.00648 (2017).
- [8] W. Ferreira, A. Vlachos, Emergent: a novel data-set for stance classification, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2016, pp. 1163–1168.
- [9] R. Assaf, M. Saheb, Dataset for arabic fake news, in: 2021 IEEE 15th International Conference on Application of Information and Communication Technologies (AICT), IEEE, 2021, pp. 1–4.
- [10] N. O'Brien, S. Latessa, G. Evangelopoulos, X. Boix, The language of fake news: Opening the black-box of deep learning based detectors (2018).
- [11] A. Bonet-Jover, A. Piad-Morffis, E. Saquete, P. Martínez-Barco, M. Á. García-Cumbreras, Exploiting discourse structure of traditional digital media to enhance automatic fake news detection, Expert Systems with Applications 169 (2021) 114340.
- [12] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, Brat: a web-based tool for nlp-assisted text annotation, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 102–107.