Semi-automatic Annotation Proposal for Increasing a Fake News Dataset in Spanish

Alba Bonet-Jover

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

Abstract

The digital era has become an ally of fake news, since it has increased the spread and amount of false information. Fake news is a global problem that causes disorder and generates fear. This phenomenon must be attacked in the same environment in which it is generated: in the digital environment. This paper presents the current state of my doctoral thesis which focuses on the linguistic modelling applied to the automatic detection of fake news through Natural Language Processing (NLP). In order to study the linguistic characteristics of fake news and to create computational models that automate its detection, labelled datasets are needed, but this is a costly task that requires time and expertise. A fake news dataset and an annotation guide were created *ad hoc* in a previous work to analyse all the parts and elements of a news item. However, after creating and training our system, we realised that the time spent was not proportional to the low annotated data obtained. The need of creating a larger corpus to train and test our hypothesis has led us to think about a way of increasing our corpus without spending so much time. For that purpose, a semi-automatic annotation is proposed for reducing time while increasing speed and quantity of the examples annotated. This proposal, besides allowing us to make progress in our research, may facilitate the creation of datasets, which are essential in NLP research.

Keywords

Natural Language Processing, Human Language Technologies, Fake news detection, Semi-automatic annotation, Corpus annotation, Corpus creation

1. Justification of the research

We are living in the global disinformation era, an era in which the excess of information is causing an infodemic, "a situation in which a lot of false information is being spread in a way that is harmful", as defined in the Collins Dictionary. Fake news has always existed and has been used for different purposes throughout history. However, the difference lies in the fact that in the current era there are more powerful dissemination tools than paper, radio or oral speeches: social media and the Internet.

The information manipulation is commonplace in the digital era and to that fact must be added the development of new technologies and the arrival of the Internet. All those factors has increased the fast spreading of fake news via social media and online digital newspapers, thereby increasing confusion and social damage. Words have a considerable power in shaping people's beliefs and opinions [1] and with the current COVID-19 pandemic, the excess of information and

Doctoral Symposium on Natural Language Processing from the PLN.net network 2021 (RED2018-102418-T), 19-20 October 2021, Baeza (Jaén), Spain.

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

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

the amount of disinformation digitally disseminated are raising unfounded fears and confusing population with hoaxes and fake news.

It is necessary to tackle this problem in the same environment in which they are created and spread: in the digital media. Due to the fast dissemination, it is impossible for humans to assume and analyse such a large amount of information in such a short period of time. For that reason, Human Language Technologies are needed to automate tasks and develop computational models in this field. The modelling of a deception language, as well as the manual annotation of news are key steps for automating the detection of fake news. To that end, labelled datasets are needed to train, but annotation is one of the most time-consuming and financially costly tasks in Natural Language Processing [2] and it requires human expertise, time and consistency. The amount of labelled corpora for research in NLP is low, even more in languages other than English, such as Spanish, a language in which there are few resources for this task.

The main objective of the thesis is to shape a deceptive language model of fake news in order not only to detect them automatically, but also to justify the decision of that detection. To that end, we manually created and annotated a fake news dataset, but to keep training our model, we need to obtain a larger corpus but also to reduce the time spent in those tasks. We propose a new Spanish dataset that is being collected and annotated combining manual and automatic processes. This semi-automatic proposal may facilitate the creation of our dataset by reducing cost and time.

This paper is structured as follows: Section 2 presents an overview of the most relevant scientific literature; Section 3 describes the new corpus, the updated version of the annotation guide and the new proposal of semi-automatic annotation; Section 4 introduces the changes that have affected the methodology in comparison with our first work and new experiments we are conducting for this research; Section 5 presents some problems encountered for discussion; Section 6 presents conclusions and future work and the paper ends with the references used for the writing of this article.

2. Background and related projects

In order to substantiate the approach of the thesis and the progress made, it is important to analyse the state of the art related to existing datasets dealing with fake news and annotation proposals available so far.

2.1. English datasets

Several datasets have been published in English to develop computational systems for the fake news detection.

- The first public fake news detection and fact-checking dataset was released by Vlachos and Riedel [3]. It was composed of 221 statements collected from PolitiFact[1] and Channel 4[2]. This dataset presented a five-label-tag classification: True, MostlyTrue, HalfTrue, MostlyFalse and False.

[1] http://www.politifact.com/
[2] http://blogs.channel4.com/factcheck/

- The dataset EMERGENT was presented by Ferreira and Vlachos [4]. It was created from rumour sites and Twitter accounts and labelled by journalists. It contained 300 claims and 2595 associated news articles. A stance label (for, against, observing) was assigned to the news article headline with respect to the claim and, in parallel, the veracity of the claim was established following three values (true, false, unverified).
- Wang [5] introduced the LIAR dataset, a broad dataset consisting of 12836 real-world short statements collected from PolitiFact[3] and covering several topics. It was manually labelled following a scale of six fine-grained labels: pants-fire, false, barely-true, half-true, mostly-true and true.
- Patwa et al. [6] released a fake news dataset of 10700 fake and real news focused on the COVID-19. This dataset was manually annotated according to two labels (real, fake).
- The first COVID-19 Twitter fake News dataset (CTF) was recently introduced by Paka et al. [7] consisting of a mixture of labelled and unlabelled tweets. The novelty lies in the semi-supervised attention neural model that works with unlabelled data to learn the writing style of tweets. This corpus was also manually annotated using a two-scale labels: Fake and Genuine.

2.2. Spanish datasets

In NLP, and particularly in the field of the fake news detection, corpora built in the language of Cervantes are scarce in comparison with those in the language of Shakespeare. Some interesting datasets for our research are:

- An opinion Spanish dataset consisting of statements covering 3 topics. It contained 100 true and 100 false statements for each topic, labelled and manually verified. With this corpus, Almela et al. [8] sought to find deceptive cues in written language in Spanish.
- A Spanish Fake News Corpus introduced by Posadas-Durán et al. [9] and composed of 491 true news and 480 fake news collected from online resources. For labelling the news, two values of veracity were considered: true and fake. To compile the corpus, keywords were identified by answering the questions What, Who, How, When and Where.
- A corpus built in Spanish and Italian including fake news spread in Facebook and Twitter in both countries. Two official Italian and Spanish fact-checking agencies, Maldita.es[4] and Bufale Un Tanto Al Chilo[5], analysed and classified the news as “fake”. With this dataset, Mottola [10] proposes an analysis of structural and linguistic features of fake news.

3. Description and objectives

To the authors’ knowledge and according to the literature consulted, available datasets in this domain are focused on labelling the news as a whole. Even if they usually present a scale of different degrees of truthfulness, they consider the whole text as a sole unit. In addition, some of them focus on news published on social media and cover a variety of topics.

[3] http://www.politifact.com/
[4] https://maldita.es/
[5] https://www.butac.it/

How does our proposal differ from those datasets? Our contribution focus on building a corpus in Spanish (both from Spain and Latin America), due to the lack of labelled resources in this language; focused on the health domain and COVID-19, because of the current pandemic situation; and composed of news collected from digital newspapers, in order to study the traditional news structure and content. Our dataset is manually collected and labelled according to the Inverted Pyramid and the 5W1H journalistic techniques. Our hypothesis lies in the fact that fake news mixes true and false information, so we propose a fine-grained annotation allowing to determine not only the full document veracity but also the veracity of each essential content element and structure parts of a news item. Notwithstanding these metrics, our proposal is adaptable to any domain and language.

In our previous work, a Spanish fake news dataset, called FNDeep Dataset, focused on the health domain and composed of 200 news (95 fake news and 105 true news) was built to support our hypothesis that knowing the veracity value separately can help to detect the overall veracity of a news item [11]. News was manually search and collected from several online newspapers or blogs and the information was checked in official fact-checking agencies belonging to the International Fact Checking Network[6]. Furthermore, we introduced an *ad hoc* annotation guide (FNDeep Annotation Scheme) and conducted several experiments to train our architecture, consisting of two main layers (Structure and Veracity). In our published paper it is shown that determining the veracity of each structural element and each 5W1H component separately influences the global veracity of a news item. However, the results, although they are good, show the need to train with a larger number of examples.

As stated in Section 1, annotating a corpus manually requires a large investment of time and effort, which makes the process slow and the number of labelled data small. To solve this problem, we are working on a combined proposal that allows to obtain and annotate news automatically, but at the same time to be checked by a human expert. For testing our proposal, we need to build another dataset to compare the corpus created by a semi-automatic process with the corpus created entirely manually (our previous corpus). As our first dataset and the annotation guide are described in detail in [11] and in [12], this work presents the new dataset, which is being created from scratch to compare it with our first dataset, and the new proposal of semi-automatic annotation.

3.1. Creation of a new fake news dataset

The aim is to check whether the change of approach and the assistance of a semi-automatic annotation allows to increase the speed and the number of annotated examples. To this end, we need to build a new corpus combining manual and semi-automatic approaches. The new corpus we are building keeps the topic (health and Covid-19), the language (Spanish) and the structure (Inverted Pyramid). However, to make this new dataset more accurate, length and format have been better defined. Regarding length, this dataset will contain news presenting a similar number of paragraphs, so that the time of annotation can be calculated on the basis of texts with similar length. With respect to the format, posts, guides, FAQs or social media posts are omitted and only news items presenting the traditional journalistic format are being

[6]https://www.poynter.org/ifcn/

collected. News following a specific format to refute a claim (that is, news with fact-check format) is also being discarded for this corpus.

3.2. FNDeep Annotation Scheme V.2.: reorientation and improvement of the guide

The complexity of our annotation lies in annotating all the content of a news item related to the Inverted Pyramid and the 5W1H journalistic techniques. The first one consists in splitting the structure of a news item into five common parts: Headline, Subtitle, Lead, Body and Conclusion. Those parts meet the Inverted Pyramid technique by placing the most relevant information at the beginning of the news item and the least relevant at the end. One of the improvements made in this regard is the annotation of the Body part. The other concept used, called 5W1H, lies in obtaining the answers to six key questions (Who, What, When, Where, Why and How) that allow to communicate a story in a complete, accurate way. The annotation of these content elements remains the same, but it has been redirected and well-defined by adding some semantic relations.

However, the most important change made with respect to our first corpus is the classification system that indicates the veracity value of each part/element. We have adopted a reliability rating instead of using a truthfulness rating, that is we have replaced the veracity attributes of True, Fake and Unknown by the attributes Reliable and Unreliable. This new classification is more accurate, since the classification of true and false depends on extra-textual factors (reader, context, external knowledge), whereas a classification of reliable or unreliable is based on a purely textual and linguistic analysis, which allows to obtain an analysis prior to the fact-checking task. Besides the reliability rating, other attributes have been added to some tags to mark semantic relations and add additional information to our annotation.

- TITLE-STANCE: this attribute, only used in the **Headline**, indicates whether or not the information presented in the Body is consistent with the information of the headline. This consistency is represented by the following values: Agree, Disagree or Unrelated.
- MAIN-EVENT: this attribute, only used with the **What** tag, allows to mark the main event of the story and it helps to differentiate it with other secondary events. A news item could contain more than one "main event".
- RELATED: content elements (5W1H) corresponding to the same event are linked with this semantic relation to differentiate multiple events appearing in the same passage. A sentence could contain more than one event, and each event may include its own 5W1H. This attribute is used by connecting all the content tags to the What tag.
- ROLE: this attribute is only used in the **Who** tag to indicate the role played by the subject/entity of the event. This function can be indicated with one of these three values: Subject (the Who causes the event), Target (the Who receives the effects of the event), or Both (when the Who performs both functions).
- ELEMENTS-OF-INTEREST: other tags allowing to create a more accurate report and to train more features are those related to style, ambiguity, lack of data, exaggeration, key terminology and phraseology, and orthotypography.

3.3. Semi-automatic annotation proposal

As stated at the beginning of this paper, the annotation task is essential in NLP since it allows to train computational models and to automate many human tasks. However, “data collection is one of the challenges of conducting deception research due to the scarce availability of such datasets” [13]. There is a lack of resources to study fake news detection in languages other than English and that scarcity is due to the fact that annotating data is a costly task: it requires time, human expertise and consistency. Posadas-Durán et al. [9] states that "annotated corpora can help to increase the performance of automatic methods aiming at detecting this kind of news” and human intervention is necessary for ensuring consistency in the annotation of texts and checking the decisions made by the machine about the features learned.

After assessing the time and level of difficulty required during the creation of our first corpus, we realised that the effort and time spent in searching and annotating news were not proportional to the number of labelled texts obtained. To progress in our research, a larger dataset is needed to train as many examples as possible but only with manual tasks it is difficult to quickly increase a corpus. Our aim lies in increasing speed and data while reducing time and effort. A semi-automatic annotation may automatically select the news and propose an automatic annotation based on our scheme.

Human intervention remains important, as the expert may check that the news set selected meets the needs of the corpus and may confirm or deny the annotation proposal. In this way, the expert ensures consistency and accuracy of the dataset while saving time in collecting and annotating news. The expert may not spend so much time searching for news items or annotating news from scratch. The assisted system may allow to create a larger, updated dataset more quickly. For carrying out this proposal, active learning techniques will be used to assist the annotation process, as this technique increases the performance of the learning model while reducing the amount of annotated data required [14].

4. Methodology and experiments

Our aim is to compare efficacy, speed and accuracy between manual annotation and assisted annotation. For that, some relevant changes were made concerning the methodology of the creation and annotation of the dataset. It is important to highlight that the annotator and the general methodology applied for this work are the same as for our first corpus. However, the semi-automatic proposal has led to change the way of collecting and classifying news, as well as the annotation tool.

4.1. Methodology

With regard to the compilation, instead of manually collect the news, the system trained with the annotated data proposes a selection of interesting news to be trained.

Regarding the annotation tool, we have chosen the Brat tool for this new corpus because it is an intuitive annotation tool allowing to annotate quickly, accurately and easily. In addition, it presents a visual and comfortable interface that facilitates the annotation task.

Last but not least, the verification process is being modified. For our first corpus, all news was manually chosen and verified in an official fact-checking agency. However, with the implementation of the first version of the assisted annotation recommendation system, the system randomly selects news, regardless of whether they have been verified by an agency or not. At that point we realised that a change of approach was needed. The analysis should focus on the textual and linguistic level to know whether a piece of a news item is reliable or not, not whether it is true or fake, since for that classification we need external knowledge.

4.2. Experiments

In order to test this combined proposal of manual annotation supported by an automatic system, three experiments will be carried out. The objective is to compare the assisted annotation with the manual annotation. To this end, each experiment will train the same number of news, which will be annotated by the same expert annotator. However, each experiment will use different approaches, from fully manual creation and annotation tasks to progressive automation of both tasks. This proposal is still being refined and tested.

EXPERIMENT 1: 30 news items will be manually searched by the expert annotator, chosen according to certain metrics (format, language, topic, length, etc.) and manually annotated without assistance, following the annotation scheme created *ad hoc*. With this first experiment, we want to calculate the time it takes the expert to perform both tasks manually, without the help of the system.

EXPERIMENT 2: 30 news items will be automatically selected by the system after having trained with the news of the first experiment, checked by the expert annotator and manually annotated. The aim here is to compare if the automatic selection of news helps to save time, as the expert only has to check whether the selection meets the requirements of the corpus and to manually annotate them.

EXPERIMENT 3: 30 news items will be pre-annotated by the semi-automatic system, and the annotator will correct and verify the annotation assisted by the system. This experiment may help to know if the semi-automation of both tasks allows to faster increase the corpus.

5. Specific elements of the research for discussion

We are still working on developing our semi-automatic annotation proposal. However, there is an issue that needs to be further analysed: the difficulty to automatically find fake news. Fact-checking agencies are constantly fighting against fake news and that continuous pressure makes fake news hard to find. Sometimes, news with a clearly format of fake news (containing spelling mistakes, capital letters, full stops, alarmist and offensive messages, etc.) disappears or is removed. This difficulty is increased when sources publish both fake and true content, since a source cannot be exactly classified and the system can propose news "considered unreliable" that are actually reliable and that can lead to an unbalanced corpus.

Another difficulty encountered is that there are cases where disinformation is spread in social media (Facebook, Twitter, WhatsApp), in the form of subjective or alarmists posts or through chain messages. We focus on disinformation in form of traditional digital news, but many times this disinformation is published only in social media.

The change of perspective has helped us to refine the classification of news. Our project focus on being a support to the fact-checking task, a previous stage. Our objective is to offer a preliminary report that allows to get a preconceived idea of a news item, to justify whether it is reliable or not, so that at a later stage it can be verified by fact-checking techniques. The veracity of a news item (true/fake) cannot be determined by language alone; external knowledge is needed to verify the information. The new classification into reliable or unreliable is more accurate and novel.

6. Conclusions and future work

This paper presents the current state of my doctoral thesis which focus on modelling a deceptive language applied to the automatic detection of fake news. Fake news has become a global problem that is damaging society in several ways: it causes fears, prejudices, hate and insecurity. Fake news makes us vulnerable. For fighting against this problem, we seek to model a deceptive language of fake news by studying the structure and content of news separately to predict its global veracity value. For that purpose, labelled data is needed. We have already created and trained a fake news dataset for our research, but a larger dataset is needed to keep training our system. As our first corpus was entirely collected and annotated manually, we need to create a new corpus combining both manual and automatic approaches to test the semi-automatic annotation.

Considering that annotated corpora are scarce in this field, specially in Spanish, and that the manual annotation of a corpus is a slow and difficult process, we propose to boost our dataset by implementing a semi-automatic annotation which may assist the expert annotator. This assisted annotation may allow to reduce time and effort while increasing speed, accuracy and labelled data. This proposal is currently being studied and tested to improve our corpus in future works and to continue combating digital disinformation.

Acknowledgments

This research work has been partially funded by Generalitat Valenciana through project “SIIA: Tecnologías del lenguaje humano para una sociedad inclusiva, igualitaria, y accesible” with grant reference PROMETEU/2018/089, by the Spanish Government through the projects RTI2018-094653-BC22: “Modelang: Modeling the behavior of digital entities by Human Language Technologies” and RTI2018-094653-B-C21: “LIVING-LANG: Living Digital Entities by Human Language Technologies”, as well as being partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER). Furthermore, I would like to thank my research team for all the work done so far: Estela Saquete, Patricio Martínez, Alejandro Piad, Suilan Estévez, Mario Nieto, Victor Belén. I also thank Miguel Ángel García Cumbreras for his participation and work in this research.

References

[1] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 conference on empirical methods in natural language processing, 2017, pp. 2931–2937.

[2] 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.

[3] A. Vlachos, S. Riedel, Fact checking: Task definition and dataset construction, in: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Association for Computational Linguistics, Baltimore, MD, USA, 2014, pp. 18–22. doi:10.3115/v1/W14-2508.

[4] 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. doi:10.18653/v1/N16-1138.

[5] W. Y. Wang, "liar, liar pants on fire": A new benchmark dataset for fake news detection, CoRR abs/1705.00648 (2017).

[6] 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, arXiv preprint arXiv:2011.03327 (2020).

[7] W. S. Paka, R. Bansal, A. Kaushik, S. Sengupta, T. Chakraborty, Cross-sean: A cross-stitch semi-supervised neural attention model for covid-19 fake news detection, Applied Soft Computing (2021) 107393.

[8] A. Almela, R. Valencia-García, P. Cantos, Seeing through deception: A computational approach to deceit detection in spanish written communication, Linguistic Evidence in Security, Law and Intelligence 1 (2013) 3–12.

[9] J. Posadas-Durán, H. Gomez-Adorno, G. Sidorov, J. Escobar, Detection of fake news in a new corpus for the spanish language, Journal of Intelligent and Fuzzy Systems 36 (2019) 4868–4876. doi:10.3233/JIFS-179034.

[10] S. Mottola, Las fake news como fenómeno social. análisis lingüístico y poder persuasivo de bulos en italiano y español, Discurso & Sociedad (2020) 683–706.

[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] A. Bonet-Jover, The disinformation battle: Linguistics and artificial intelligence join to beat it (2020).

[13] E. Saquete, D. Tomás, P. Moreda, P. Martínez-Barco, M. Palomar, Fighting post-truth using natural language processing: A review and open challenges, Expert systems with applications 141 (2020) 112943.

[14] M. Kholghi, L. Sitbon, G. Zuccon, A. Nguyen, Active learning: a step towards automating medical concept extraction, Journal of the American Medical Informatics Association 23 (2016) 289–296.