Review of: "Automatic Content Analysis Systems: Detecting Disinformation in Social Networks"

José Ramón Méndez Reboredo[1]

1 University of Vigo

Potential competing interests: No potential competing interests to declare.

The manuscript is centered on the detection of fake news (disinformation, incompleteness, propaganda, fake news...). The authors provide a state-of-the-art review and discuss the need for using popular large language models such as ChatGPT4 (at least in the abstract). The authors use a corpus of 20,000 articles (50% fake, 50% normal) to conduct some experimentation and observe the connection between sentiment analysis and fake news. They also test a model using logistic regression.

I find the target problem very difficult to solve. In most cases, the identification of fake news will require access to real information. For instance, let's think about the following possible situation:
+ "50 positions for jobs in the public sector are being offered. 1000 candidates apply, but only 10 of these 50 positions are occupied."
This is an undesirable situation in an administration because it implies that the selection board did not conduct a proper selection process or that the education systems are not effective. So news can hide information about the number of positions offered or the number of candidates:
+ "10 new public employees have been assigned last week"....
What happened here? I was hiding information. How can I find this information as "propaganda," "disinformation," or "fake news"? In order to detect this, I will need to access the whole real/original information. Even for a human, it is difficult to detect that some information has been hidden. One can think that there were only 10 positions offered and all of them were assigned (which is not true). Summarizing, the problem is really hard, and I believe that solving it adequately requires combining external information (different sources that can be used to contrast information, access to the real information published in some sources,....).

I find the problem of fake reviews identification is more similar to fake news. There are some reviews about that. You can read some articles such as those with DOIs 10.1109/ACCESS.2021.3075573, 10.1007/s10618-021-00772-6, 10.1109/CCAA.2018.8777594

I disagree with the sentence "The problem of detecting fake sources of information can be similar to the tasks of detecting spam, especially when using statistical methods of machine learning." Certainly, similar techniques have been used to solve both problems until now, which seems not to be optimal. In fact, there is a great difference between fake news and spam. Spam usually has an explicit commercial purpose; there is a URL where you can buy a product, a price, ... Fake news does not really contain an explicit explanation of their goal, does not expose the real information, ... These are two really different problems. So I believe they should be addressed by using different methods. The sentence aims to think that it is a good idea to use the same techniques for both problems. To detect spam, I can find a URL in the message, I can find words (or named entities) with the meaning of "price" (named entities of the type currency)... But to detect fake news, we would probably need to develop techniques to compare the target message with other sources of information to evaluate if the information included in the target message is complete. I would expect in the manuscript argumentation in this line.

The sentence "Machine learning techniques used to classify text such as tweets or emails to determine whether it is spam are already successfully used to detect spam." seems not to be correct. Maybe "Machine learning techniques used to classify texts, such as tweets or emails, to determine whether they are spam or not, have been also used to detect fake news."

The sentence "Once the features are defined, they can be used for classification using various machine-learning techniques that involve training with a teacher.". With a teacher? ML techniques are trained with data.

Authors say "The task of disinformation detection is similar to the task of spam detection, as both aim to separate genuine texts from fake or false ones. However, it is important to consider these tasks separately due to several differences.". But differences are never explained. These differences should be explained to understand the inner details of the target problem.

Char n-grams are most commonly used in language detection than in spam filtering. I believe they have low relevance both in spam filtering and fake news detection.

When talking about Word2Vec, I would talk about word embedding. I would include in this group GloVe, fastText, BERT, ELMo,...

In addition to BoW (Bag Of Words), there are also some representations of texts using specific features for each text (Fdez-Riverola....). Texts are represented in indexing data structures (memory-based). The purpose of these representations is to find similar texts to the target one and use a voting scheme to classify text. Instance Retrieval Networks, used by S. J. Delany, Cunningham..., or Enhanced Instance Retrieval Networks (Fdez-Riverola, ...), I believe, can be mentioned.

I do not like using unordered lists in scientific articles. Such language constructions give the impression of a schematic rather than a scientific text.

The state of the art described is too similar to the one that would be depicted for spam detection. But remember that those problems are radically different. For me, the only idea that fits the identification of fake news is "Also, analyzing comments and comparing them with similar articles can be a useful method to assess the veracity of news or text. If the majority of such articles do not corroborate the data in the news, this may indicate that the news may be biased or fake." Also, from page 8, there are specific solutions for fake news identification.

The state of the art seems not to have a coherent order. I would arrange the methods explained in a coherent manner and explain in the first paragraph of the section the criteria used to introduce the methods.

I really appreciate figures 1-6 showing source code and results. However, I believe code is unnecessary and that the relevant information can be presented in a graphically better (diagrams, charts, ...).

Please specify in a formal manner the preprocessing steps carried out on texts. The sentence "For both sub-datasets, basic text clean-up procedures were performed, such as changing text to lowercase, removing punctuation, cleaning up location and author tags, and removing stop words, etc." contains "etc.," which indicates there are additional steps that are not documented.

The experimental protocol only uses logistic regression, BOW, and sentiment analysis. However, the State of the Art has a lot of algorithms and possible representations that are not included in the experimental protocol. There is no comparison with other similar texts to try to contrast the information. It seems that the authors used some techniques such as the ones used in spam detection (which is a problem very different). I find the experimentation does not reflect the potential of the techniques included at the end of the state of the art (more connected with fake news identification). There is no experimentation using GPT4 (in the abstract, authors talk about its importance). For me, the experimentation included in the manuscript is not useful and is not connected with the goal of the manuscript.

I recommend that authors include a subsection containing a formal explanation of the experimental protocol and then present the results achieved through the experimental protocol.

The last three paragraphs included in the conclusion section are, in fact, a numbered list. I recommend not using this kind of format in formal scientific texts.

If authors indicate the relevance of GPT4 in the detection of fake news, I believe this idea should be reflected also in the conclusions.

I would describe future work directions at the end of the conclusions section.

I find the manuscript should be greatly improved following the advice included in this revision and encourage authors to do it.