Review of: "A Comparative Analysis of Advertising in the 2020 Presidential Elections & Phoenix Mayoral Elections using Natural Language Processing"

Edgar León-Sandoval[1]

1 Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM)

I read this paper with interest and high expectations, for the idea is interesting and worth the publication and time invested in it. However, this paper lacks rigor in pretty much every area; only the problem statement is engaging, for the rest is a data science tutorial (most of the references posted are indeed links to tutorials), and overall, it is not suitable as an academic submission.

Regarding form:

References

- Some are missing.
- Inconsistent formatting
- Inconsistent use of URL links

In the project, the paper "Going Micro to Go Negative? Targeting Toxicity using Facebook and Instagram Ads" has been used by us.

- Reference?
- Written language needs work

NLP techniques are the best tools for answering the question.

- What question?

Facebook uses its data from advertisements to support our question. The Facebook Ads library API provides reliable data sources from year to year and every election campaign. Although the dataset still has some empty values or outlines, we can still remove them with data cleaning technology. Based on this, we think the dataset can support and answer our question very well.

- What question?
- Reliable data sources? As in more than one?
- Data cleaning technology?

- Is this really enough to say that the dataset is appropriate?

"We intended to recreate the model used by the original study that studied …"

- Grammar, again
- Intended? As in "we tried to, but couldn't"?
- Are the authors sharing this model?

The same source researchers have primarily used to collect their dataset. The researchers have also used financial sponsorships different contestants receive and the different advertisers' information for data analysis.

- What researchers?
- Source missing?

Sentiment analysis of the political campaign advertisements is being performed to compare the toxicity of advertisements.

- Awesome!

A reason for our case to not require an iterative approach is that the model has a small dataset. This will lead to the model being improperly trained and could be overfitted.

- A mitigation strategy here would be nice
- "Will lead to"… well, did it?

The section "**Data Cleaning**" refers to some considerations in data collection, but no cleaning is mentioned.

The section "Data Security" needs more explanations and a grammar check.

The section "Process and Methodology" starts with actual data cleaning. These common procedures should be included in the appendix, or just relegated to a simple explanation, and any interested party can check the details with the shared code.

- Same thing with the screenshots of the Python code utilized, and listed code.
- Sharing the code is nice, but does it have to be here in this tutorial form?
- "For the RNN, we have used the powerful PyCharm as an IDE"—how is this relevant?
- No code is shown in the data gathering portion.

In the section "Data Modeling," authors should refrain from the overly use of "very."

This section reads as a simple, often overly simple, textbook. Are we learning about recurrent networks in this text?

An example of over-simplicity in the text:

> *Overfitting*

> *There are many parameters in RNN, so it is prone to overfitting problems.*

The section "Data Analysis" is very short and lacks data, results, graphs, or any representation of the results of all the code shown earlier in the work.

The section "Conclusion" is written in such a way as to refer to previously shown examinations and results. Neither was presented.

The references list seems short and is not properly formatted. It should include DOIs, not links to data science blogs or other tutorials.