Reviews of 1109 - *"A Benchmark Dataset of Check-worthy Factual Claims"*

Reviewer 4 (SPC/Associate Editor)

Meta-Review and Roadmap

The paper proposes a novel dataset on US political debates. The authors collected all the US presidential debates from 1960 to today. They performed an annotation through crowdsourcing, classifying each statement into three labels: check-worthy factual, unimportant factual and non-factual. Experts were also employed to rate the job of the crowdsourced annotators.
All reviewers appreciate the paper and its contributions.

For the final submission, the authors should:
• add inter-coder agreement coefficients such as Krippendorff's alpha or Cohen's kappa.
• Share (if possible) all the collected labels from separated annotators
• Fix the link to the web-based platform
• Some additional details of users and parties of politicians should be added.

Reviewer 1 (reviewer/PC member)

Overall evaluation

+2 = accept

Reviewer's Confidence

3 = Fairly confident that I've adequately considered all aspects

Reviewer's methodological expertise

3 = Knowledgeable: Knowledgeable in this methodological approach

Paper summary

The paper proposes a novel dataset on US political debates. The authors collected all the US presidential debates from 1960 to today. They performed an annotation through crowdsourcing, classifying each statement into three labels: check-worthy factual,

unimportant factual and non-factual. Experts were also employed to rate the job of the crowdsourced annotators.

Reasons to accept

- Fully comprehensive dataset over all US presidential debates
- Very sound annotation process, interface and quality evaluation

Reasons to reject

- The dataset does not provide all the collected labels from separated annotators, irrespectively from the agreement
- Not enough analysis of debatable statements

Comments for authors

Overall, I believe the dataset will be very impactful, and extensively used for future research projects. There are a couple of relatively minor issues, but nothing particularly important to recommend further major revision instead of acceptance.

Please see detailed comments in the next review sections.

Originality of work

The work is somehow incremental, since it is an extension of half (the Check-worthy part) of Nakov et. al. (2018). This is not an issue however, since a more comprehensive and precise dataset is needed for further research.

Potential impact of results

Analysis of political data is a very hot topic, with a lot of research going on, and the dataset will be for sure analysed a lot and the submission cited.

Quality of execution

It is mostly good. The annotation process is very sound, the annotation interface and process well-designed, and it can eventually recycled for further annotations on the same data, or other related tasks. The amount of data collected is also very extensive, at least considering only US data. It would be nice, at least to the majority of people who are not US national, to extend the annotation and analysis to other countries and other languages.

I have however spotted a potential issue regarding debatable statements with respect to the annotation labels. The authors make an interesting error analysis regarding

mislabelled samples, and state that they consider the UFS<->{CFS,NFS} errors less serious than CFS<->NFS. I may however guess this is not necessarily correct, since I would assume that a confusing NFS (or UFS) statement may still contain a somehow hidden or confusing factual claim behind, maybe difficult to see. Therefore, I would in turn assume that errors of the form NFS->UFS->CFS (false positive) to be less important than a CFS->UFS->NFS (false negative), where a factual statement would be discarded, especially when there is a strong disagreement in the annotations. This is however an issue that can be somehow fixed by providing all the raw annotation labels, not only when there is an agreement, to allow eventually the end user to make its own analysis. It might be however worth to consider for future data collection and annotation rounds.

Quality of presentation

The paper is very clear, and easy to read and replicate.

Adequancy of citations

All the important papers seem correctly cited.

Ethical concerns (if any)

None to report.

Reviewer 2 (reviewer/PC member)

Overall evaluation

+2 = accept

Reviewer's Confidence

3 = Fairly confident that I've adequately considered all aspects

Reviewer's methodological expertise

3 = Knowledgeable: Knowledgeable in this methodological approach

Paper summary

This paper introduces ClaimBuster --- a dataset for benchmarking of check-worthy factual claims from all U.S. presidential debates. The dataset consists of about 23,5K sentences belonging to three categories, i.e., non-factual, unimportant factual, and check-worthy factual statements.

The sentences have been extracted from a raw political diabetes dataset which is also publicly available. After that, the authors used a self-implemented web-based platform for labeling the sentences. The labeling was done by paid participants aware of U.S politics. The participants could assign one of the three categories for a randomly selected sentence. Skipping hard sentences and requesting more context was possible.

I think this dataset would be of much interest to many researchers and I recommend it for acceptance.

Reasons to accept

- compared to other existing datasets, ClaimBuster has two advantages: (i) size and (ii) ideologically-free annotation strategy.
- two possible use cases for the dataset that would allow for the development models for automatic detection of check-worthy claims and of factual claims.

Reasons to reject

- a bit unfortunate is that the demographics of the recruited participants are rather briefly described. A better description would have been nice to see and would have helped to clearly distinguish this dataset with respect to political/ideological biases of fact-checkers of the two datasets the authors described in the related work section.

Comments for authors

The authors provided some descriptive insights into the dataset and the labeling quality. I would have appreciated seeing precalculated inter-coder agreement coefficients such as Krippendorff’s alpha or Cohen’s kappa.

Unfortunately, the link to the web-based platform was not working at the time of reviewing.

Originality of work

-

Potential impact of results

The presented dataset addresses an important issue for the ICWSM community and can help to solve pressing misinformation problems in online political discourses, e.g., in social media platforms.

Quality of execution

The quality of the labels was ensured by the initial platform and task training of the participants, and on-site workshops. The labeling performance of the participants was assessed by using screening sentences stemming from a ground-truth dataset compiled by experts. The authors implemented a stopping condition to ensure that each of the remaining sentences received a reasonable number of labels form top-performing participants.

Quality of presentation

The structure of the datasets is clear. The authors also used speaking names for files and feature sets. Sentiment score calculated using the Alchemy API has been added to ease the reproducibility of previously published results using this dataset.

FAIR principles are being followed. The authors made the dataset publicly available and citable using Zenodo. The reusability of the dataset is assured by a ReadMe-File accompanying the data. The dataset consists of CSV files that can be easily transformed in other formats. It was nice to see that one of the authors provided his ORCID to the data publishing platform.

Adequancy of citations

ok

Ethical concerns (if any)

-

Reviewer 3 (reviewer/PC member)

Overall evaluation

+2 = accept

Reviewer's Confidence

3 = Fairly confident that I've adequately considered all aspects

Reviewer's methodological expertise

3 = Knowledgeable: Knowledgeable in this methodological approach

Paper summary

The paper presents factchecked results by taking content of US general election presidential debates. The data is carefully annotated by around 100 people which makes

this data unique and useful. A platform is presented to users for performing the annotation tasks, monitoring payments, ect.

Reasons to accept

+ Dataset is quite unique, large and carefully examined.
+ Paper is well-written, explains all possible reasons, stats and related work.
+ The general trends of debate and online Transcripts data is also helpful

Reasons to reject

Following are some concerns-
- The study is missing a baseline comparison. How well are these users versus other automatic tools? At least one recent NLP model should have been employed to compare.
- What are demographics of users? Are there returning users and they are somewhat contributing more among 100 annotators?
- What about comparison of the party? Are left or right using more a certain type of sentences?

Comments for authors

- An NLP model to test 'Check-worthy Factual Sentence' or not can be from Miriam Redi et al. "Citation Needed: A Taxonomy and Algorithmic Assessment of Wikipedia's Verifiability" WWW 2019 which is used by Wikipedia.
- Some additional details of users and party of politicians can be added.

Originality of work

High: The dataset is a great contribution.

Potential impact of results

High: The paper and dataset can be used for multiple research work. It can be helpful for an inter-disciplinary study.

Quality of execution

- Good: Clearly explains the motivation and research approach.

Quality of presentation

Good: Well written, justified claims and clear examples.

Adequancy of citations

Good: Few recent are missing.

Ethical concerns (if any)

(blank)

Return to submission and reviews