CIKM 2015
The 24th ACM International Conference on Information and Knowledge Management
19-23 October 2015, Melbourne, Australia

Reviews For Paper

Track	Knowledge Management
Paper ID	1138
Title	Detecting Check-worthy Factual Claims in Presidential Debates

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
Is the submission relevant to the KM track?	Yes
What do you think of the ideas in the submission?	Some novelty
Is the writing clear?	Yes
Overall recommendation	+3: Should accept
How to improve the submission (for the camera-ready, if accepted or for resubmission to another venue, if rejected)? List your three suggestions.	See detailed review
Detailed review	The poster considers the problem of automatically detecting factual claims, which, e.g., can be passed over to journalists for further investigation. The authors evaluate several classifiers with different attributes to extract the sentences corresponding to important claims that need to be checked. Page 3, 1st paragraph: The discussion concerning the derivation of ground truth is a bit fuzzy. Did you use at all the scores of participants that were not ranked as top-quality? How were disagreements handled in that case. There were 7.4K sentences labeled by 2 top-quality participants but the participants scores agreed in only 1.5K of these sentences. This is lower than 1/3rd, which you would achieve in the case of completely random scoring. Did you have a look why the participants had such a high disagreement?

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	

Is the submission relevant to the KM track?	Yes
What do you think of the ideas in the submission?	Some novelty
Is the writing clear?	Yes
Overall recommendation	+3: Should accept
How to improve the submission (for the camera-ready, if accepted or for resubmission to another venue, if rejected)? List your three suggestions.	There are a couple of text overflows, most notably in both columns of the first page. The patterns used in Figure 4 are difficult to discern when the page is printed (either because of printer artifacts or because of the small size). It would make the data much easier to read if you made them more distinct. In Table 2, it might be helpful to delimit rows better so that, for example, it would be easy to identify all of the "NBC" algorithm results.
Detailed review	I think you have a good model, and that you used a good quality (and sized) dataset. Your ground-truth collection seems sound, and you obviously spent a good deal of time on making sure that it was done correctly. The tangential observation of the increase in factual claims in recent years is, in and of itself, quite interesting. You seem to have identified a nice variety of features, and your comparison matrix of the top results seems logical. I would like to see a bit more information as to model growth and flexibility. For example, in 2016, it is highly probable that the rhetoric will include far more references to the NSA, privacy, and "ISIS" than the debates from 1960. Do you consider this to be a problem? A cursory application to other media sources would have been nice as well, but I understand that you faced space limitations.

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
Is the submission relevant to the KM track?	Yes
What do you think of the ideas in the submission?	Incremental
Is the writing clear?	Yes
Overall recommendation	+3: Should accept
	1. This paper adopts supervised classification methods to detect sentences

How to improve the submission (for the camera-ready, if accepted or for resubmission to another venue, if rejected)? List your three suggestions.	that are check-worthy factual claims, in the scenario of presidential debates. I think there are plenty of improvements that the authors can apply to further increase the precision and recall. For example, using n-gram instead of uni-gram, using stemming and removing stop words (some features shown in Figure 6 are stop words) for pre-processing, etc. Also I wonder why feature selection didn't further improve the performance, maybe it will be different after preprocessing and using n-grams. 2. Check-worthy factual statement detection is a very popular and meaningful problem to study. I think it would be more impactful if the authors can apply their learned classifier on social media data to extract and analyze large scale of check-worthy factual posts related to presidential debates. 3. I think the authors need to have a more concrete definition or understanding on ``check-worthy''. Currently according to their explanation in section 2, it is a very subjective criteria for human annotators to label. I also wonder what is the inter rater reliability of the labeling.
Detailed review	1. Since all the sentences are generated from transcripts of presidential debates, I wonder if the detection should be done on a statement level (multiple sentences about one factual statement) instead of a sentence level. 2. In their cross validation, I wonder how the performance will be if the authors use the 2004 and 2008 debates as training and the 2012 debates as testing. I think we need to verify if the trained classifier can be used on a new dataset without learning any of its content and topic relevant features. 3. Existed works have shown that some content features (such as keywords, RTs, question marks, urls, etc.) are useful for very similar detection/classification tasks, such as information credibility detection [2] and controversial statement detection. (some related work: Zhe Zhao, Paul Resnick, Qiaozhu Mei, Enquiry Minds: Early Detection of Rumors in Social Media from Enquiry Posts, WWW 2015). I wonder if the authors have considered these features in building their classifier.