Detecting Check-worthy Factual Claims in Presidential Debates

Naeemul Hassan
Department of Computer Science and Engineering
University of Texas at Arlington
naeemul.hassan@mavs.uta.edu

Chengkai Li
Department of Computer Science and Engineering
University of Texas at Arlington
cli@uta.edu

Mark Tremayne
Department of Communication
University of Texas at Arlington
tremayne@uta.edu

ABSTRACT

Public figures such as politicians make claims about "facts" all the time. Journalists and citizens spend a good amount of time checking the veracity of such claims. Toward automatic fact checking, we developed tools to find check-worthy factual claims from natural language sentences. Specifically, we prepared a U.S. presidential debate dataset and built classification models to distinguish check-worthy factual claims from non-factual claims and unimportant factual claims. We also identified the most-effective features based on their impact on the classification models' accuracy.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; H.4.m [**Information Systems Applications**]: Miscellaneous; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms

Algorithms, Design, Experimentation

Keywords

computational journalism; fact checking; text classification

1. INTRODUCTION

Public figures such as politicians make claims about "facts" all the time. Oftentimes there are false, exaggerated and misleading claims on important topics, due to careless mistakes and even deliberate manipulation of information. With technology and modern day media helping spread information to mass audiences through all types of channels, there is a pressing need for checking the veracity of factual claims important to the public. Journalists and citizens spend good amount of time doing that. More and more dedicated platforms and institutes are being created for fact checking. According to a census from the Duke University Reporters' Lab,[1] the number of fact checking platforms such as *PolitiFact.com* and *FactCheckEU.org* has increased from 59 (May 2014) to 89 (January 2015), a *50.8%* increase in eight months. This genre of investigative reporting has become a basic feature of political coverage, especially during elections, and plays an important role in improving political discourse and increasing democratic accountability [8, 5].

The process of fact checking requires many challenging steps—extracting natural language sentences from speeches, interviews, press releases, campaign brochures and social media; separating factual claims from opinions, beliefs, hyperboles, questions, and so on; detecting topics of factual claims and discerning which are "check-worthy"; assessing the veracity of such claims, which itself requires collecting information and data, interviewing experts, and presenting evidence and explanations.[2]

Part of the goal of *computational journalism* [3, 4] is use computing to automate fact checking [11, 9]. A fully automatic fact checking system is not yet within our reach. It calls for breakthroughs in several fronts related to the aforementioned fact checking steps. This paper's focus is on detecting check-worthy factual claims from natural language sentences, specifically transcripts of presidential debates.

We model this problem as a classification task and we follow a supervised learning approach to tackle it. We constructed a labeled dataset of spoken sentences by presidential candidates during 2004, 2008 and 2012 presidential debates. (Data collection for earlier debates is still in progress.) Each sentence is given one of three possible labels—it is not a factual claim; it is an unimportant factual claim; it is an important factual claim. We trained and tested several multi-class classification models using the labeled dataset. Experiment results demonstrated promising accuracy of the models. We further identified and analyzed the most-effective features in the models.

We envision, during presidential debates of U.S. Election 2016, for every sentence spoken by the two candidates and extracted into transcripts, our model will immediately predict whether the sentence has a factual claim and whether checking its truthfulness is important to the public. Furthermore, factual claims will be ranked by their significance, which will help professional and citizen journalists focus on

CIKM'15, October 19–23, 2015, Melbourne, VIC, Australia.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806652.

[1] http://reporterslab.org/fact-checking-census-finds-growth-around-world/

[2] http://www.politifact.com/truth-o-meter/article/2013/nov/01/principles-politifact-punditfact-and-truth-o-meter/

the right target. Although so far we have only collected data related to presidential debates, the studied models can be possibly applied on other types of text, including speeches, radio/TV interviews, and social media.

To the best of our knowledge, no prior study has focused on computational methods for detecting factual claims and discerning their importance. The most relevant line of work is subjectivity analysis of text (e.g., [12, 1, 10]) which classifies sentences into objective and subjective ones. However, not all objective sentences are check-worthy important factual claims. Wu et al. [11] studied how to model the quality of facts and find their supporting arguments and counterarguments. Vlachos and Riedel [9] analyzed the tasks in fact checking and presented a dataset of factual claims collected from *PolitiFact.com* and *Channel4.com*. Another area of related research is checking information credibility in microblog platforms. For instance, [13] finds trending rumors containing disputed factual claims. [2, 6] focus on assigning credibility scores to tweets. The scoring models are highly dependent on Twitter-specific features such as the credibility of twitter users. A tweet with high credibility does not necessarily contain a check-worthy factual claims.

2. PROBLEM FORMULATION

We categorize sentences in presidential debates into three categories. Below, each category is explained with examples.

Non-Factual Sentence (*NFS*): Subjective sentences (opinions, beliefs, declarations) and many questions fall under this category. These sentences do not contain any factual claim. Below are some examples.

- *But I think it's time to talk about the future.*
- *You remember the last time you said that?*

Unimportant Factual Sentence (*UFS*): These are factual claims but not check-worthy. In other words, the general public will not be interested in knowing whether these sentences are true or false. Fact-checkers do not find these sentences as important for checking. Some examples are as follows.

- *Next Tuesday is Election Day.*
- *Two days ago we ate lunch at a restaurant.*

Check-worthy Factual Sentence (*CFS*): These sentences contain factual claims and the general public will be interested in knowing whether the claims are true or false. Journalists look for these type of claims for fact checking. Some examples are:

- *He voted against the first Gulf War.*
- *Over a million and a quarter Americans are HIV-positive.*

Our goal is to automatically detect *CFS*s. We model it as a supervised learning problem. Specifically, we model it as a multi-class classification problem where the classes are *NFS*, *UFS* and *CFS*.

3. DATA COLLECTION

In order to construct a dataset for developing and evaluating approaches to detect check-worthy factual claims, we used presidential debate transcripts. The first general election presidential debate was held in 1960. Since then, there have been 14 elections till 2012. In 1964, 1968 and 1972, no presidential debate was held. There were 2 to 4 debate episodes in each of the remaining 11 elections. A total of 30 debate episodes spanned 1960–2012. We parsed the debate transcripts and identified the speaker for each sentence. There are a total of 123 speakers including 18 presidential candidates, moderators and guests. The whole dataset consists of 28029 sentences. We are interested in sentences spoken by the presidential candidates only. There are 23075 such sentences. We discarded very short sentences (less than 5 words long) and we were remained with 20788 sentences. Figure 1 shows the distribution of sentences among 30 debate episodes. Figure 2 shows the average length of sentences. These figures show that recent candidates used more sentences and shorter ones than earlier candidates.

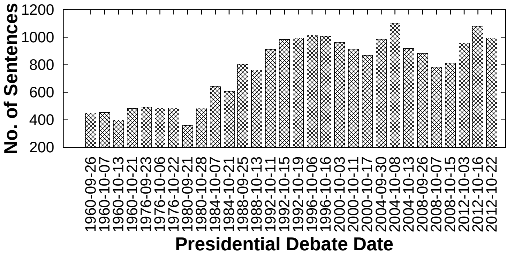


Figure 1: Sentence Distribution

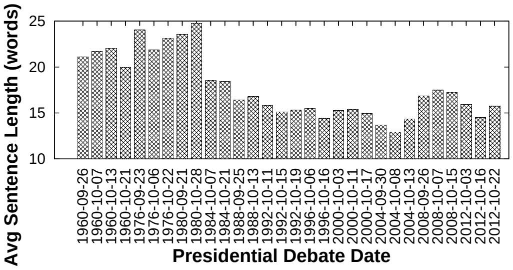
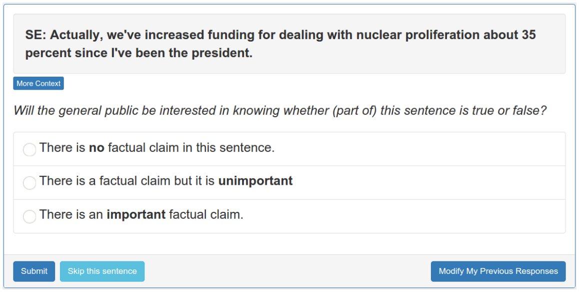


Figure 2: Average Sentence Length in Words

To label the sentences, we developed a data collection website. Journalists, professors and university students were invited to participate in the survey. There was a reward system to encourage high quality answers. A participant was given one sentence at a time and was asked to label it with one of the three possible options as shown in Figure 3. If the participant was not sure about their answer, they could click the "More Context" button to see five preceding sentences of the given sentence. They could also click the "Skip" button to skip the sentence.



SE: Actually, we've increased funding for dealing with nuclear proliferation about 35 percent since I've been the president.

Will the general public be interested in knowing whether (part of) this sentence is true or false?

○ There is **no** factual claim in this sentence.

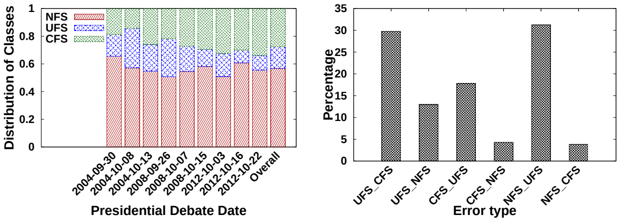
○ There is a factual claim but it is **unimportant**

○ There is an **important** factual claim.

Figure 3: Data Collection Interface

In 15 days, we accumulated 140 participants. To detect spammers and low-quality participants, we used 123 screening sentences (48 *NFS*s, 32 *UFS*s and 43 *CFS*s). These sentences were picked from all debate episodes. Three domain experts agreed upon their labels. On average, one out of every ten sentences given to a participant (without letting the participant know) was randomly chosen to be a screening question and selected from the pool of 123 sentences. The participants were scored in the range of [0.0, 1.0] based on their performance on the screening sentences. Those scored more than 0.85 were considered top-quality participants.

We aimed to get the latest debates labeled first. Sentences from one debate episode were randomly presented to the participants. Once all sentences in an episode were labeled by at least two participants, we moved on to the next episode. The data collection is still in progress. So far, 2012, 2008 and 2004 presidential debates (12 debate episodes) have been labeled. For training and evaluating our classification models, we only used a sentence if its label was agreed upon by two top-quality participants. Thereby we got 1571 sentences (882 *NFS*s, 252 *UFS*s, 437 *CFS*s). Figure 4a shows the distribution of these sentences' class labels. One interesting observation is that recent presidential candidates were making more check-worthy factual claims than earlier candidates.



(a) Class Distribution (b) Error Distribution

Figure 4

Category	Type	# of Features	Example
Sentiment	continuous	1	-0.5, 0.0, 0.5
Length	discrete	1	5, 10, 50
Word (W)	continuous	6130	*debt*, *five*
POS Tag (P)	discrete	43	*VBD*, *CD*
Entity Type (ET)	discrete	26	*Person*

Table 1: Summary of Feature Categories

4. FEATURE EXTRACTION

We extracted multiple categories of features from the sentences. Table 1 summarizes these features. We use the following sentence to explain the features.

When President Bush came into office, we had a budget surplus and the national debt was a little over five trillion.

Sentiment: We used AlchemyAPI[3] to calculate a sentiment score for each sentence. The score ranges from -1 (most negative sentiment) to 1 (most positive sentiment). The above sentence has a sentiment score -0.846376.

Length: This is the word count of a sentence. Natural language toolkit NLTK[4] was used for tokenizing a sentence into words. The example sentence has length 21.

Word (*W*): We used words in sentences to build *tf-idf* features. After discarding rare words that appear in less than three sentences, we got 6130 words. We did not apply stemming or stopword removal.

Parts of Speech (POS) Tag (*P*): We applied NLTK POS tagger on all sentences. There are 43 POS tags in the corpus. We constructed a feature for each tag. For a sentence, the count of words belonging to a POS tag is the value of the corresponding feature. In the example sentence, there are 3 words (*came*, *had*, *was*) with POS tag *VBD (Verb, Past Tense)* and 2 words (*five*, *trillion*) with POS tag *CD (Cardinal Number)*.

Entity Type (*ET*): We used AlchemyAPI to extract entities from the sentences. There are 2727 entities in the labeled sentences. These entities belong to 26 types. The above sentence has an entity *"Bush"* of type *"Person"*. We constructed a feature for each entity type. For a sentence, its number of entities of a particular type is the value of the corresponding feature.

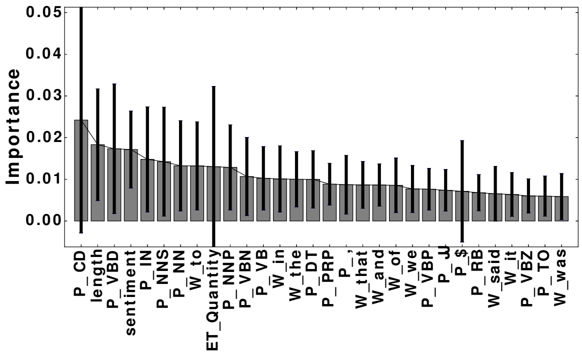


Figure 5: Feature Importance

Feature Selection: There are 6201 features in total. To avoid over-fitting and to attain a simpler model, we performed feature selection. We trained a random forest classifier for which we used GINI index to measure the importance of features in constructing each decision tree. The overall importance of a feature is its average importance over all the trees. Figure 5 shows the importance of the 30 best features in the forest. The black solid lines indicate the standard deviations of importance values. Category types are prefixes to feature names. It is unsurprising that *P_CD* is the top discriminator—check-worthy factual claims are more likely to contain numeric values (*45%* of *CFS* sentences in our dataset contain numeric values) and non-factual sentences are less likely to contain numeric values (*6%* of *NFS* sentences in our dataset contain numeric values). Figure 6 shows the value distributions across all three classes for the four most important features. It depicts the features' discriminative capacities.

5. CLASSIFICATION

We performed 4-fold cross-validation using several supervised learning methods, including Multinomial Naive Bayes Classifier (*NBC*), Support Vector Classifier (*SVM*) and Random Forest Classifier (*RFC*). Table 2 shows these classifiers' performance in terms of precision (p), recall (r), f-measure (f) and Cohen's kappa coefficient (κ). We experimented with four combinations of features—Word (*W*), Word + POS Tag (*W_P*), Word + POS Tag + Entity Type (*W_P_ET*), and the 100 most important features (*best_100*). *Sentiment* and *Length* were included in all the combinations. The *SVM* classifier paired with *W_P* achieved *70%*, *72%* and *70%* weighted average precision, recall and f-measure, respectively. *RFC* and *SVM* outperformed *NBC* in most cases. To understand the level of agreement between classifiers and human participants, we used κ coefficient. According to the guideline set in [7], *RFC* and *SVM* agreed moderately with the participants and *NBC* agreed fairly.

All the classification models had better accuracy on *NFS*s and *CFS*s than *UFS*s. This is not surprising, since *UFS* is

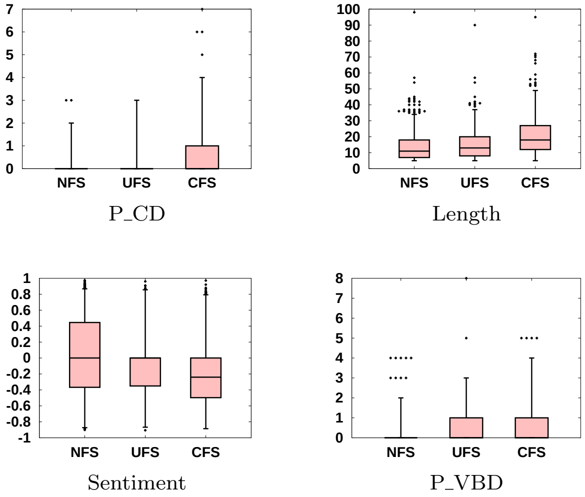
[3] http://www.alchemyapi.com/

[4] http://www.nltk.org/

algorithm	features	p_NFS	p_UFS	p_CFS	p_wavg	r_NFS	r_UFS	r_CFS	r_wavg	f_NFS	f_UFS	f_CFS	f_wavg	κ
NBC	W	0.55	0.00	0.60	0.47	1.00	0.00	0.01	0.55	0.71	0.00	0.02	0.39	0.01
SVM	W	0.75	0.48	0.67	0.69	0.88	0.20	0.64	0.70	0.81	0.27	0.64	0.68	0.45
RFC	W	0.66	0.60	0.85	0.71	0.97	0.03	0.47	0.69	0.78	0.06	0.61	0.62	0.35
NBC	W_P	0.65	0.00	0.79	0.58	0.98	0.00	0.44	0.67	0.78	0.00	0.56	0.59	0.32
SVM	W_P	0.76	0.45	0.69	0.70	0.89	0.22	0.65	0.72	0.82	0.29	0.67	0.70	0.48
RFC	W_P	0.70	0.72	0.73	0.71	0.95	0.04	0.56	0.70	0.81	0.08	0.63	0.64	0.40
NBC	W_P_ET	0.69	0.00	0.77	0.61	0.98	0.00	0.51	0.70	0.81	0.00	0.61	0.63	0.38
SVM	W_P_ET	0.74	0.47	0.70	0.69	0.90	0.23	0.62	0.71	0.81	0.31	0.66	0.69	0.47
RFC	W_P_ET	0.70	0.57	0.77	0.70	0.97	0.04	0.56	0.71	0.81	0.08	0.65	0.65	0.41
NBC	best_100	0.74	0.31	0.67	0.65	0.88	0.21	0.52	0.67	0.80	0.25	0.58	0.66	0.40
SVM	best_100	0.72	0.43	0.76	0.69	0.92	0.13	0.56	0.70	0.80	0.16	0.63	0.66	0.42
RFC	best_100	0.74	0.56	0.68	0.70	0.91	0.14	0.64	0.72	0.82	0.22	0.66	0.68	0.45

Table 2: Comparison of NBC, SVM and RFC coupled with various feature sets, in terms of Precision (p), Recall (r), F-measure (f) and Cohen's kappa coefficient (κ). *wavg* denotes weighted average of corresponding measure across three classes.

between the other two classes and thus the most ambiguous. The top-quality participants faced screening sentences 1395 times and made incorrect judgment 208 times (*14.9%*). Figure 4b shows the percentages of different error types among these 208 cases. For instance, *UFS_CFS* represents the cases in which *UFS*s were incorrectly labeled as *CFS*s by participants. It is evident from this figure that even the top-quality participants made more mistakes when class *UFS* is in question.



P_CD

Length

Sentiment

P_VBD

Figure 6: Value Distributions of the Four Most Important Features

6. CONCLUSIONS AND FUTURE WORK

We presented a supervised learning based approach to automatically detect check-worthy factual claims from presidential debate transcripts. We conducted a closely monitored survey to collect labels on sentences from the debates. We performed feature extraction and important feature selection. Preliminary experiment results show that the models achieved *85%* precision and *65%* recall in classifying check-worthy factual claims. We plan to carry on future research along the following directions:

- We will complete label collection for the remaining (1960–2000) debate transcripts. We will analyze how classification performance changes by training data from different years' debates. For the upcoming 2016 U.S. presidential election, we will offer a website that ranks check-worthy factual claims, which can assist journalists and citizens in prioritizing their fact checking endeavor.
- We will extend the study to other types of texts, including interviews, congressional records, and social media.
- We aim at improving feature extraction, feature selection, and classification methods, to obtain better classification accuracy. We also plan to develop methods for tackling claims spanning over multiple sentences.

Acknowledgments The authors have been partially supported by NSF grants IIS-1018865, CCF-1117369 and IIS-1408928. Any opinions, findings, and conclusions in this publication are those of the authors and do not necessarily reflect the views of the funding agencies.

7. REFERENCES

[1] P. Biyani, S. Bhatia, C. Caragea, and P. Mitra. Using non-lexical features for identifying factual and opinionative threads in online forums. *Knowledge-Based Systems*, 69:170–178, 2014.

[2] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW*, pages 675–684, 2011.

[3] S. Cohen, J. T. Hamilton, and F. Turner. Computational journalism. *CACM*, 54(10):66–71, Oct. 2011.

[4] S. Cohen, C. Li, J. Yang, and C. Yu. Computational journalism: A call to arms to database researchers. In *CIDR*, pages 148–151, 2011.

[5] L. Graves. *Deciding What's True: Fact-Checking Journalism and the New Ecology of News.* PhD thesis, COLUMBIA UNIVERSITY, 2013.

[6] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *Social Informatics*, pages 228–243. 2014.

[7] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

[8] B. Nyhan and J. Reifler. The effect of fact-checking on elites: A field experiment on us state legislators. *American Journal of Political Science*, 59(3):628–640, 2015.

[9] A. Vlachos and S. Riedel. Fact checking: Task definition and dataset construction. In *ACL*, pages 18–22, 2014.

[10] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *CICLing*, pages 486–497. 2005.

[11] Y. Wu, P. K. Agarwal, C. Li, J. Yang, and C. Yu. Toward computational fact-checking. In *PVLDB*, pages 589–600, 2014.

[12] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *EMNLP*, pages 129–136, 2003.

[13] Z. Zhao, P. Resnick, and Q. Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *WWW*, pages 1395–1405, 2015.