Detecting Check-worthy Factual Claims in Presidential Debates

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 develop classification models 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 accuracy of the classification models.

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 has increased from 59 (May 2014) to 89 (January 2015), a *50.8%* increase in eight months. Among such platforms the popular ones include *PolitiFact.com*, *FactCheckEU.org* and *FullFact.org*. 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 textual/audial sources such as 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 the "check-worthy" claims; 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 labelled 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 labelled 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 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 coun-

[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/

terarguments. 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 micro-blog platforms. For instance, [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 presented and illustrated with examples.

Non-Factual Sentence (*NFS*): Subjective sentences including opinion, belief, declaration and questions fall under this category. These sentences do not contain any factual claim. Some examples are-

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

Unimportant Factual Sentence (*UFS*): These are factual claims but not check-worthy. In other words, it is possible to check veracity of these claims but general public will not be interested in knowing whether these sentences are true or false. So, journalist or reporters do not find these sentences as important for checking. Some examples are-

- *I am a graduate of the U.S.*
- *Next Tuesday is Election Day.*
- *Two days ago we ate lunch at a restaurant.*

Check-worthy Factual Sentence (*CFS*): These are the sentences where factual claims are present and 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.*
- *There are 1.4 million children without health insurance.*
- *Over a million and a quarter Americans are HIV-positive.*

Our goal is to automatically detect check-worthy factual claims or *CFS*. We model the problem as a supervised learning problem. Specifically, we model this as a multiclass classification problem where the classes are *NFS*, *UFS* and *CFS*.

In the following sections, we explain the data collection, feature extraction and supervised model building process.

3. DATA COLLECTION

In order to construct a dataset for developing and evaluating approaches to detect check-worthy factual claims, we use 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 spanning 1960–2012. We parse the debate transcripts and identify the speaker for each sentence. There are a total of 123 speakers including 18 presidential candidates and others (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 were remained with 20788 sentences. Figure 1 shows distribution of sentences among these 30 debate episodes. Figure 2 shows the average length of sentences. From these two figures, it is observed that recent candidates used shorter sentences than earlier candidates.

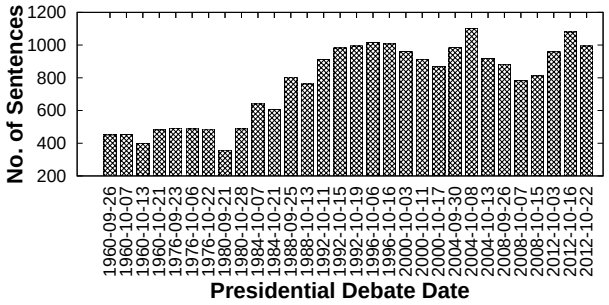

Figure 1: Sentence Distribution

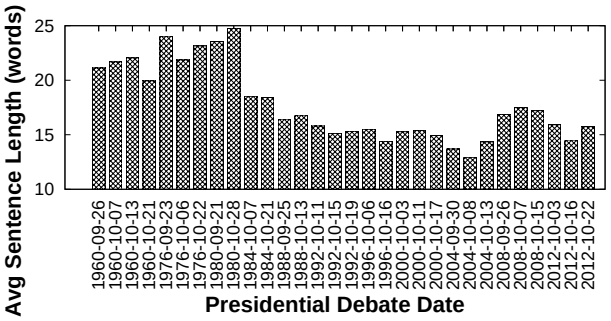

Figure 2: Average length of sentences

To collect ground-truth, we developed an online survey website.[3] Journalists, professors and university students were invited to participate in the survey. There was a reward system to encourage high quality answers. A participant is given one sentence at a time and is asked to label it with one of three possible options as shown in Figure 3. There was also an option for skipping a given sentence. A "More Context" button showed five preceding sentences of the given sentence if the participant is not sure of the context.

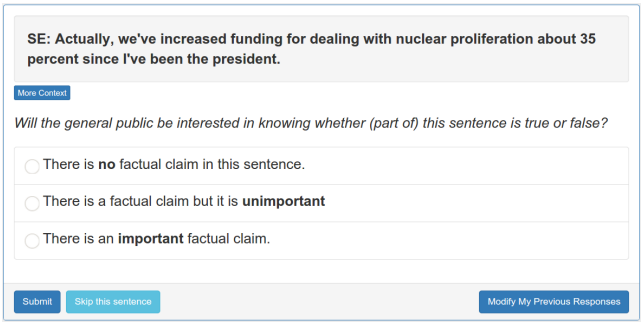

Figure 3: Data Collection Interface

In 15 days, we accumulated 140 participants. To control spammers and low-quality participants, we used 123 screening sentences (48 of *NFS* class, 32 of *UFS* class and 43 of *CFS* class). These screening sentences were picked from all debate episodes. Three domain experts agreed upon the labels of these screening sentences. For every 10 sentences labeled by a participant, there was at least one from the screening sentences. Based on performance in the screening sentences, we give a ranking score to the participants. The score ranges from [0.0–1.0]; higher ranking score means higher quality. Participants with ranking score >= 0.85 were regarded as top-quality participants.

[3]URL not provided due to conference's double blind rule

Our procedure was to get the latest debates' sentences labeled first. Sentences from one debate episode are randomly presented to the participants. Once all the sentences of an episode is labeled by at least two participants, we move on to the next episode. Currently the survey is in progress. So far, ground-truth collection for 2012, 2008 and 2004 presidential debates (12 debate episodes) is completed. We have 7465 sentences labeled by two participants. For quality purposes, we only use those sentences as ground-truth for which labels were agreed upon by two top-quality participants. This way we have 1571 sentences (NFS = 882, UFS = 252, CFS = 437) in the ground-truth set. Figure 4 shows distribution of the class labels in the debate episodes. One interesting thing to observe from this figure is, recent presidential candidates are making more check-worthy factual claims than earlier presidential candidates.

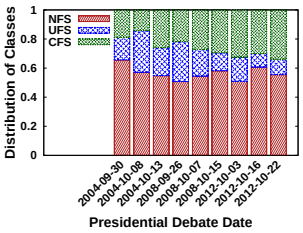

Figure 4: Class Distribution

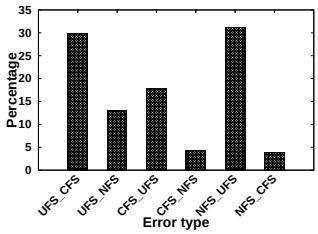

Figure 5: Error Distribution

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

Table 1: Summary of Feature categories

4. FEATURE EXTRACTION

We extract the following categories of features from the sentences. Each category is explained with the example sentence below. Table 1 gives a summary of the categories.

Example: *In the last four years, there have been twice as many bankruptcies as new jobs created.*

Sentiment: We use third-party library AlchemyAPI[4] to calculate sentiment score of each sentence. The score ranges from -1 to 1. A score close to -1 means the sentence is of negative sentiment. A positive score indicates the sentence is of positive sentiment. For example, the above sentence has sentiment score -0.729107.

Length: This is the word count of a sentence. Natural language toolkit NLTK[5] is used for tokenizing a sentence into words. The example sentence has length 16.

Word (*W*): We use the words in sentences to build *tf-idf* features. We discard very rare words (words present in less than three sentences). After removal, we have 6130 words.

Parts of Speech (POS) Tag (*P*): We use NLTK POS tagger to find POS tags of all the words in a sentence. In the example sentence, the words *"four"*, *"the"* and *"jobs"* have POS tags *Cardinal Number (CD)*, *Determiner (DT)* and *Plural Noun (NNS)*, respectively. There are 43 POS tags in the whole corpus. We build a feature for each tag. Count of words belonging to a POS tag in a particular sentence forms the value of the corresponding feature.

Entity Type (*ET*): We use AlchemyAPI for entity extraction from the sentences. There are 2727 entities. These entities are organized into 26 types. As an example, the above sentence has an entity *"four years"* of type *"Quantity"*. We build a feature for each entity type. Number of entities of a particular type in a sentence determines the feature value of that entity type for that sentence.

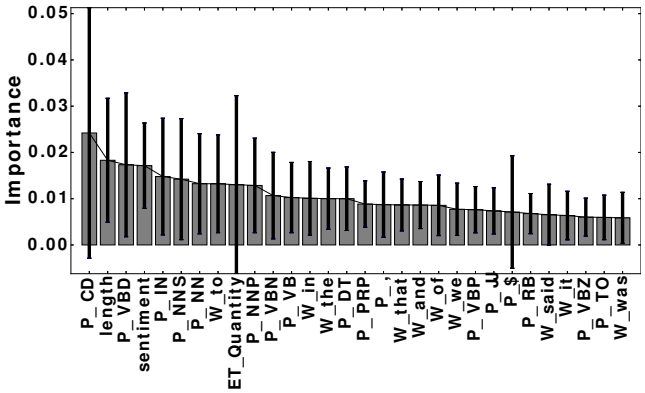

Figure 6: Feature Importance

Feature Selection: There are 6201 features in total. To avoid over-fitting and attain a simpler model, we perform best feature selection. We train a Random Forest classifier. It uses GINI index to calculate importance of each feature. Figure 6 shows importance of best 30 features along with their inter-trees variability. Notation of category types are appended as a suffix to the feature names. It is unsurprising that *P_CD* is the top discriminator. Generally check-worthy factual claims are more likely to contain numeric values (*45%* of *CFS* sentences in our dataset contain numeric value) and non-factual sentences are less likely to contain numeric values (*6%* of *NFS* sentences in our dataset contain numeric value). Figure 7 shows box plots of four most important features. It depicts discriminative capacity of the features.

5. CLASSIFICATION

We have experimented with various supervised learning methods including Multinomial Naive Bayes Classifier (*NBC*), Support Vector Classifier (*SVM*) and Random Forest Classifier (*RFC*). 4-fold cross-validation was performed for parameter tuning. The ground-truth set was divided into training and test set having 3:1 ratio.

Table 2 shows performance of these classifiers in terms of precision (p), recall (r), f-measure (f) and Cohen's κ (kappa) coefficient. We have studied four combinations of features– Words (*W*), Words + POS Tags (*W_P*), Words + POS Tags + Entity Types (*W_P_ET*) and 100 most important features (*best_100*). *Sentiment* and *Length* were included in all the combinations. The *SVM* classifier paired with *W_P* achieves *70%*, *72%* and *70%* weighted average precision, recall and f-measure, respectively. We observe that *RFC* and *SVM* outperform *NBC*. In some cases, (*W_P*) for example, *NBC* classified all the test instances as either *NFS* or *CFS*. That is why we observe zero recall and precision in corresponding table cells. To understand the level of agreement between a classifier and the ground-truth, we used κ coefficient. According to the guideline set in [7], *RFC* and *SVM* agreed moderately with the ground-truth and *NBC* agreed fairly.

[4]http://www.alchemyapi.com/
[5]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 learning models with respect to various feature set in terms of Precision (p), Recall (r), F-measure (f) and κ value. *wavg* denotes weighted average of corresponding measure across three classes.

All the classification models show better performance in classifying *NFS* and *CFS* sentences than classifying *UFS* sentences. It may be due to the intrinsic difficulty in identifying sentences of *UFS* class. The top-quality participants faced screening sentences totally 1395 times and made error judgment in 208 cases (*14.9%*). Figure 5 shows percentages of different error types in these 208 cases. *UFS_CFS* means the correct label of the screening sentence is *UFS* but the participant erroneously labeled it as *CFS*. It is evident from this figure that even the top-quality participants made more mistakes when class *UFS* is in question.

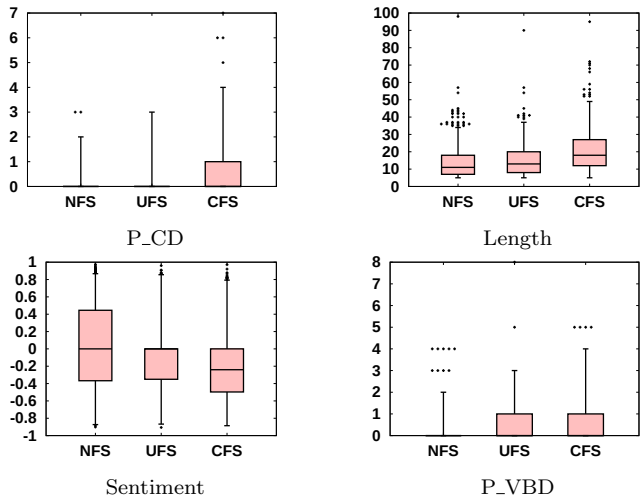

Figure 7: Box plots of four most important features

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 ground-truth 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 ground-truth 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, speeches, 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.

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, 2014.

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

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

[4] S. Cohen, C. Li, J. Yang, and C. Yu. Computational journalism: A call to arms to database researchers. In *CIDR*, 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: A real-time web-based system for assessing credibility of content on twitter. *CoRR*, abs/1405.5490, 2014.

[7] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 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*, 2014.

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

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

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

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