Automated Fact-Checking of Claims in Argumentative Parliamentary Debates

Nona Naderi
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada
nona@cs.toronto.edu

Graeme Hirst
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada
gh@cs.toronto.edu

Abstract

We present an automated approach to distinguish true, false, stretch, and dodge statements in questions and answers in the Canadian Parliament. We leverage the truthfulness annotations of a U.S. fact-checking corpus by training a neural net model and incorporating the prediction probabilities into our models. We find that in concert with other linguistic features, these probabilities can improve the multi-class classification results. We further show that dodge statements can be detected with an F_1 measure as high as 82.57% in binary classification settings.

1 Introduction

Governments and parliaments that are selected and chosen by citizens' votes have *ipso facto* attracted a certain level of trust. However, governments and parliamentarians use combinations of true statements, false statements, and exaggerations in strategic ways to question other parties' trustworthiness and to thereby create distrust towards them while gaining credibility for themselves. Creating distrust and alienation may be achieved by using ad hominem arguments or by raising questions about someone's character and honesty (Walton, 2005). For example, consider the claims made within the following question that was asked in the Canadian Parliament:

Example 1.1 [Dominic LeBlanc, 2013-10-21] *The RCMP and Mike Duffy's lawyer have shown us that the Prime Minister has not been honest about this scandal. When will he come clean and stop hiding his own role in this scandal?*

These claims, including the presupposition of the second sentence that the Prime Minister has a role in the scandal that he is hiding, may be true, false, or simply exaggerations. In order to be able to analyze how these claims serve their presenter's purpose or intention, we need to determine their truth.

Here, we will examine the linguistic characteristics of true statements, false statements, dodges, and stretches in argumentative parliamentary statements. We examine whether falsehoods told by members of parliament can be identified with previously proposed approaches and we find that while some of these approaches improve the classification, identifying falsehoods by members of parliament remains challenging.

2 Related work

Vlachos and Riedel (2014) proposed to use data from fact-checking websites, such as PolitiFact for the fact-checking task and suggested that one way to approach this task would be using the semantic similarity between statements. Hassan et al. (2015) used presidential debates and proposed three labels — *Non-Factual*, *Unimportant Factual*, and *Check-worthy Factual* sentence — for the fact-checking task. They used a traditional feature-based method and trained their models using sentiment scores using AlchemyAPI, word counts of a sentence, bag of words, part-of-speech tags, and entity types to classify the debates into these three labels. They found that the part-of-speech tag of cardinal numbers was the most informative feature and word counts was the second most informative feature. They also found that check-worthy actual claims were more likely to contain numeric values and non-factual sentences were less likely to contain numeric values.

Patwari et al. (2017) used primary debates and presidential debates for analyzing check-worthy statements. They used topics extracted using LDA, entity history and type counts, part-of-speech tuples, counts of part-of-speech tags, unigrams, sentiment, and token counts for their classi-

Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 60–65
Brussels, Belgium, November 1, 2018. ©2018 Association for Computational Linguistics

Label	True	False	Dodge	Stretch	Total
#	255	60	70	93	478

Table 1: Distribution of labels in the *Toronto Star* dataset

Label	#
True	1,780
Mostly true	2,003
Half true	2,152
Mostly false	1,717
False	1,964
Pants-on-fire false	867
Total	10,483

Table 2: Distribution of labels in the PolitiFact dataset

fication task. Ma et al. (2017) used a kernel-based model to detect rumors in tweets. Wang (2017) used the statements from PolitiFact and the 6-point scale of truthfulness; he compared the performance of multiple classifiers and reported some improvement by using metadata related to the person making the statements.

Rashkin et al. (2017) examined the effectiveness of LIWC (Linguistic Inquiry and Word Count) and stylistic lexicon features in determining the reliability of the news corpus and truthfulness of the PolitiFact dataset. The only reliability measurement reported on the PolitiFact dataset is by Wang (2017), who manually analyzed 200 statements from PolitiFact and reached an agreement of 0.82 using Cohen's kappa measurement with the journalists' labels. Jaradat et al. (2018) used a set of linguistic features to rank check-worthy claims. Throne et al. (2018) created a dataset for claim verification. This dataset consists of 185,445 claims verified against Wikipedia pages. Here, we do not consider any external resources and we focus only on the text of claims to determine whether we can classify claims as true, false, dodge, or stretch.

3 Data

For our analysis, we extracted our data from a project by the *Toronto Star* newspaper.[1] The *Star* reporters[2] fact-checked and annotated questions and answers from the Oral Question Period of the Canadian Parliament (over five days in April and May 2018). Oral Question Period is a session of 45 minutes in which the Opposition and Government backbenchers ask questions of ministers of the government, and the ministers must respond. The reporters annotated all assertions within both the questions and the answers as either *true*, *false, stretch,* (half true), or *dodge* (not actually answering the question). Further, they provided a narrative justification for the assignment of each label (we do not use that data here). Here is an example of the annotated data (not including the justifications):

Example 3.1 ***Q.*** [Michelle Rempel] *Mr. Speaker, [social programs across Canada are under severe strain due to tens of thousands of unplanned immigrants illegally crossing into Canada from the United States.]*$_{\textbf{\textit{False}}}$ *[Forty per cent in Toronto's homeless shelters are recent asylum claimants.]*$_{\textbf{\textit{True}}}$ *[This, food bank usage, and unemployment rates show that many new asylum claimants are not having successful integration experiences.]*$_{\textbf{\textit{False}}}$

A. [Ahmed Hussen (Minister of Immigration, Refugees and Citizenship)] *Mr. Speaker, we commend the City of Toronto, as well as the Province of Ontario, the Province of Quebec, and all Canadians, on their generosity toward newcomers. That is something this country is proud of, and we will always be proud of our tradition. [In terms of asylum processing, making sure that there are minimal impacts on provincial social services, we have provided $74 million to make sure that the Immigration and Refugee Board does its work so that legitimate claimants can move on with their lives and those who do not have legitimate claims can be removed from Canada.]*$_{\textbf{\textit{True}}}$

Here is an example of dodge annotation:

Example 3.2 ***Q.*** [Jacques Gourde] . . . *How much money does that represent for the families that will be affected by the sexist carbon tax over a one-year period?*

A. [Catherine McKenna (Minister of Environment and Climate Change)] *[Mr. Speaker, I am quite surprised to hear them say they are concerned about sexism. That is the party that closed 12 out of 16 Status of Women Canada offices.]*$_{\textbf{\textit{Dodge}}}$ *We know that we must take action*

[1] http://projects.thestar.com/question-period/index.html. All the data is publicly available.

[2] Bruce Campion-Smith, Brendan Kennedy, Marco Chown Oved, Alex Ballingall, Alex Boutilier, and Tonda MacCharles.

Features	F_1	Accuracy	Dodge	True	False	Stretch
Majority class (True)	–	53.35				
BOW (tf-idf)	49.20	53.14	55.20	67.00	4.60	24.80
+ POS	52.92	58.15	62.40	71.00	4.80	27.40
+ NUM	53.40	58.58	**63.80**	70.80	4.80	28.80
+ Superlatives (Rashkin et al., 2017)	54.24	59.42	**63.80**	71.60	9.20	30.00
+ PolitiFact predictions	**55.10**	**59.63**	63.60	**71.60**	12.80	**30.80**
BOW + NE	50.66	53.33	57.40	66.40	**17.20**	24.40

Table 3: Five-fold cross-validation results (F_1 and % accuracy) of four-way classification of fact-checking for the overall dataset and F_1 for each class.

on climate change. Canadians know that we have a plan, but they are not so sure if the Conservatives do.

For our analysis, we extracted the annotated span of the text with its associated label. The distribution of the labels in this dataset is shown in Table 1. This is a skewed dataset with more than half of the statements annotated as *true*.

We also use a publicly available dataset from PolitiFact, a website at which statements by American politicians and officials are annotated with a 6-point scale of truthfulness.[3] The distribution of labels in this data is shown in Table 2. We examine PolitiFact data to determine whether these annotations can help the classification of the *Toronto Star* annotations.

4 Method

We formulate the analysis as a multi-class classification task; given a statement, we identify whether the statement is true, false, stretch, or a dodge.

We first examine the effective features used for identifying deceptive texts in the prior literature.

- Tuples of words and their part-of-speech tags (unigrams and bigrams weighted by *tf-idf*, represented by POS in the result tables).
- Number of words in the statement (Hassan et al., 2015; Patwari et al., 2017).
- Named entity type counts, including organizations and locations (Patwari et al., 2017) (represented by NE in the result tables).
- Total number of numbers in the text, e.g., ***six** organizations heard the assistant deputy minister* (Hassan et al., 2015) (represented by NUM in the result tables).
- LIWC (Tausczik and Pennebaker, 2010) features (Rashkin et al., 2017).
- Five lexicons of intensifying words from Wiktionary: superlatives, comparatives, action adverbs, manner adverbs, modal adverbs (Rashkin et al., 2017).

In addition, we leverage the American PolitiFact data to fact-check the Canadian Parliamentary questions and answers by training a Gated Recurrent Unit classifier (GRU) (Cho et al., 2014) on this data. We will use the truthfulness predictions of this classifier — the probabilities of the 6-point-scale labels — as additional features for our SVM classifier (using the scikit-learn package (Pedregosa et al., 2011)). For training the GRU classifier, we initialized the word representations using the publicly available GloVe pre-trained 100-dimension word embeddings (Pennington et al., 2014)[4], and restricted the vocabulary to the 5,000 most-frequent words and a sequence length of 300. We added a dropout of 0.6 after the embedding layer and a dropout layer of 0.8 before the final sigmoid unit layer. The model was trained with categorical cross-entropy with the Adam optimizer (Kingma and Ba, 2014) for 10 epochs and batch size of 64. We used 10% of the data for validation, with the model achieving an average F_1 measure of 31.44% on this data.

5 Results and discussion

We approach the fact-checking of the statements as a multi-class classification task. Our baselines

[3] The dataset has been made available by Hannah Rashkin at https://homes.cs.washington.edu/~hrashkin/factcheck.html.

[4] https://nlp.stanford.edu/projects/glove/

Features	Dodge	Stretch	False
True			
Majority class	54.84	52.25	58.62
BOW	76.09	54.21	58.20
BOW + NE	75.65	52.99	61.67
BOW + LIWC	52.38	49.11	53.41
BOW + PolitiFact	**77.96**	**55.73**	58.11
BOW + NE + Politifact	76.25	53.76	**63.69**
BOW + POS + NUM + Superlative + PolitiFact	77.51	54.96	55.24
False			
Majority class	53.85	**60.00**	
BOW	81.36	55.89	
BOW + NE	**82.57**	56.91	
BOW + LIWC	52.02	53.31	
BOW + PolitiFact	80.69	52.97	
BOW + NE + Politifact	82.52	55.08	
BOW + POS + NUM + Superlative + PolitiFact	78.29	54.82	
Stretch			
Majority class	57.06		
BOW	75.15		
BOW + NE	76.93		
BOW + LIWC	45.37		
BOW + PolitiFact	79.39		
BOW + NE + Politifact	77.73		
BOW + POS + NUM + Superlative + PolitiFact	**80.59**		

Table 4: Average F_1 of different models for two-way classification of fact-checking (five-fold cross-validation).

are the majority class (truths) and an SVM classifier trained with unigrams extracted from the annotated spans of texts (weighted by *tf-idf*). We performed five-fold cross-validation. Table 3 reports the results on the multi-class classification task with these baselines and with the additional features described in section 4, including the truthfulness predictions of the GRU classifier trained on PolitiFact data. The best result is achieved using unigrams, POS tags, total number of numbers (NUM), superlatives, and the GRU's truthfulness predictions (PolitiFact predictions). We examined all five lexicons from Wiktionary provided by Rashkin et al. (2017); however, only superlatives affected the performance of the classifier, so we report only the results using superlatives.

We also report in Table 3 the average F_1 measure for classification of four labels in multi-class classification using five-fold cross-validation. The truthfulness predictions did not improve the classification of the *dodge* and *true* labels in multi-class classification setting. Superlatives slightly improved the classification of all labels except *dodge*.

We further perform pairwise classification (one-versus-one) for all possible pairs of labels to get better insight into the impact of the features and characteristics of labels.

Therefore, we created three rather balanced datasets of truths and falsehoods by randomly re-sampling the *true* statements without replacement (85 *true* statements in each dataset). The same method was used for comparing *true* labels with *dodge* and *stretch* labels, i.e., we created three relatively balanced datasets for analyzing *true* and *dodge* labels and three datasets for analyzing *true* and *stretch* labels. This allows us to compare the prior work on the 6-point scale truthfulness labels on the U.S. data with the Canadian 4-point scale.

Table 4 presents the classification results using five-fold cross-validation with an SVM classifier. The reported F_1 measure is the average of the results on all three datasets for each pairwise setting. *Dodge* statements were classified more accurately than the other statements with an F_1 measure as high as 82.57%. This shows that the answers that do not provide a response to the question can be detected with relatively high confidence. The most effective features for classifying *false* against *true* and *dodge* statements were named entities.

The predictions obtained from training the GRU model on the PolitiFact annotations, on their own, were not able to distinguish *false* from *true* and *stretch* statements. However, the predictions did help in distinguishing *true* against *stretch* and *dodge* statements. None of the models were able to improve the classification of *false* against *stretch* statements over the majority baseline.

Overall, *stretch* statements were the most difficult statements to identify in the binary classification setting. This could also be due to some inconsistency in the annotation process, with *stretch* and *false* not always clearly separated. Here is an example of *stretch* in the data:

Example 5.1 [Catherine McKenna] *Carbon pricing works and it can be done while growing the economy. . . . Once again, I ask the member opposite, "What are you going to do?" [Under 10 years of the [Conservative] Harper government, you did nothing.]*$_{\textbf{\textit{Stretch}}}$

Elsewhere in the data, essentially the same claim is labelled *false*:

Example 5.2 [Justin Trudeau] *The Conservatives promised that they would also tackle environmental challenges and that they would do so by means other than carbon pricing. . . . They have no proposals, [they did nothing for 10 years.]*$_{\textbf{\textit{False}}}$

We further performed the analysis using the two predictions of *more true* and *more false* from the PolitiFact dataset; however, we didn't observe any improvements. Using the total number of words in the statements also did not improve the results.

While Rashkin et al. (2017), found that LIWC features were effective for predicting the truthfulness of the statements in PolitiFact, we did not observe any improvements in the performance of the classifier in our classification task on Canadian Parliamentary data. Furthermore, we did not observe any improvements in the classification tasks using sentiment and subjectivity features extracted using OpinionFinder (Wilson et al., 2005; Riloff et al., 2003; Riloff and Wiebe, 2003).

6 Comparison with PolitiFact dataset

In this section, we perform a direct analysis with the PolitiFact dataset. We first train a GRU model (used a sequence length of 200, other hyperparameters the same as those of the experiment described above) using 3-point scale annotations of PolitiFact (used 10% of the data for validation). We treat the top two truthful ratings (true and mostly true) as true; half true and mostly false as stretch; and the last two ratings (false and pants-on-fire false) as false. We then test the model on three annotations of true, stretch, and false from the *Toronto Star* project. The results are presented in Table 5. As the results show, none of the false statements are detected as *false* and the overall F_1 score is lower than the majority baseline.

We further train a GRU model (trained with binary cross-entropy and sequence length of 200, other hyperparameters the same as above) using 2-point scale where we treat the top three truthful ratings as true and the last three false ratings as false. We then test the model on two annotations of true and false from the *Toronto Star* project. The results are presented in Table 6; the F_1 score remains below baseline.

The Politifact dataset provided by Rashkin et al. includes a subset of direct quotes by original speakers. We further performed the 3-point scale and 2-point scale analysis using only the direct quotes. Using only the direct quotes, also shown in Tables 5 and 6, did not improve the classification performance.

	F_1	True	Stretch	False
Majority	63			
GRU (All)	40	53	29	0
GRU (DQ)	50	75	13	8

Table 5: 3-point scale comparison of the PolitiFact data and *Toronto Star* annotations. **All**: GRU model is trained with all PolitiFact data and tested on *Toronto Star* annotations. **DQ**: GRU model is trained with only direct quotes from the PolitiFact data and tested on *Toronto Star* annotations.

	F_1	True	False
Majority	81		
GRU (All)	73	84	29
GRU (DQ)	72	88	8

Table 6: 2-point scale comparison of the PolitiFact data and *Toronto Star* annotations. **All**: GRU model is trained with all PolitiFact data and tested on *Toronto Star* annotations. **DQ**: GRU model is trained with only direct quotes from the PolitiFact data and tested on *Toronto Star* annotations.

7 Conclusion

We have analyzed classification of *truths, falsehoods, dodges,* and *stretches* in the Canadian Parliament and compared it with the truthfulness classification of statements in the PolitiFact dataset. We studied whether the effective features in the prior research can help us characterize the truthfulness in Canadian Parliamentary debates and found out that while some of these features help us identify *dodge* statements with an F_1 measure as high as 82.57%, they were not very effective in identifying *false* and *stretch* statements. The truthfulness predictions obtained from training a model on annotations of American politicians' statements, when used with other features, helped slightly in distinguishing truths from other statements. In future work, we will take advantage of journalists' justifications in determining the truthfulness of the statements as relying on only linguistic features is not enough for determining falsehoods in parliament.

Acknowledgements

This research is financially supported by an Ontario Graduate Scholarship, the Natural Sciences and Engineering Research Council of Canada, and the University of Toronto. We thank the anonymous reviewers for their thoughtful comments and suggestions.

References

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259,*.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1835–1838. ACM.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. Claimrank: Detecting check-worthy claims in arabic and english. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada. Association for Computational Linguistics.

Ayush Patwari, Dan Goldwasser, and Saurabh Bagchi. 2017. TATHYA: A multi-classifier system for detecting check-worthy statements in political debates. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2259–2262. ACM.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 105–112, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Douglas Walton. 2005. *Fundamentals of critical argumentation*. Cambridge University Press.

William Yang Wang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.