Where is your Evidence: Improving Fact-checking by Justification Modeling

Tariq Alhindi and **Savvas Petridis**
Department of Computer Science
Columbia University
tariq@cs.columbia.edu
sdp2137@columbia.edu

Smaranda Muresan
Department of Computer Science
Data Science Institute
Columbia University
smara@columbia.edu

Abstract

Fact-checking is a journalistic practice that compares a claim made publicly against trusted sources of facts. Wang (2017) introduced a large dataset of validated claims from the POLITIFACT.com website (LIAR dataset), enabling the development of machine learning approaches for fact-checking. However, approaches based on this dataset have focused primarily on modeling the claim and speaker-related metadata, without considering the evidence used by humans in labeling the claims. We extend the LIAR dataset by automatically extracting the justification from the fact-checking article used by humans to label a given claim. We show that modeling the extracted justification in conjunction with the claim (and metadata) provides a significant improvement regardless of the machine learning model used (feature-based or deep learning) both in a binary classification task (true, false) and in a six-way classification task (pants on fire, false, mostly false, half true, mostly true, true).

1 Introduction

Fact-checking is the process of assessing the veracity of claims. It requires identifying evidence from trusted sources, understanding the context, and reasoning about what can be inferred from the evidence. Several organizations such as FACTCHECK.org, POLITIFACT.com and FULLFACT.org are devoted to such activities, and the final verdict can reflect varying degrees of truth (e.g., POLITIFACT labels claims as true, mostly true, half true, mostly false, false and pants on fire).

Until recently, the bottleneck for developing automatic methods for fact-checking has been the lack of large datasets for building machine learning models. Thorne and Vlachos (2018) provide a survey of current datasets and models for fact-checking (e.g., (Wang, 2017; Rashkin et al., 2017; Vlachos and Riedel, 2014; Thorne et al., 2018; Long et al., 2017; Potthast et al., 2018; Wang et al., 2018)). Wang (2017) has introduced a large dataset (LIAR) of claims from POLITIFACT, the associated metadata for each claim and the verdict (6 class labels). Most work on the LIAR dataset has focused on modeling the content of the claim (including hedging, sentiment and emotion analysis) and the speaker-related metadata (Wang, 2017; Rashkin et al., 2017; Long et al., 2017).

However, these approaches do not use the evidence and the justification provided by humans to predict the label. Extracting evidence from (trusted) sources for fact-checking or for argument mining is a difficult task (Rinott et al., 2015; Thorne et al., 2018; Baly et al., 2018). For the purpose of our paper, we rely on the fact-checking article associated with the claim. We extend the original LIAR dataset by automatically extracting the justification given by humans for labeling the claim, from the fact-checking article (Section 2). We release the extended LIAR dataset (LIAR-PLUS) to the community[1].

The main contribution of this paper is to show that modeling the extracted justification in conjunction with the claim (and metadata) provides a significant improvement regardless of the machine learning model used (feature-based or deep learning) both in a binary classification task (true, false) and in a six-way classification task (pants on fire, false, mostly false, half-true, mostly true, true) (Section 4). We provide a detailed error analysis and per-class results.

Our work complements the recent work on providing datasets and models that enable the development of an end-to-end pipeline for fact-

[1] https://github.com/Tariq60/LIAR-PLUS

checking ((Thorne et al., 2018) for English and (Baly et al., 2018) for Arabic). We are primarily concerned on showing the impact of modeling the human-provided justification for predicting the veracity of a claim. In addition, our task aims to capture the varying degrees of truth that some claims might have and that are usually labeled as such by professionals (rather than binary true vs. false labels).

2 Dataset

The LIAR dataset introduced by (Wang, 2017) consists of 12,836 short statements taken from POLITIFACT and labeled by humans for truthfulness, subject, context/venue, speaker, state, party, and prior history. For truthfulness, the LIAR dataset has six labels: pants-fire, false, mostly-false, half-true, mostly-true, and true. These six label sets are relatively balanced in size. The statements were collected from a variety of broadcasting mediums, like TV interviews, speeches, tweets, debates, and they cover a broad range of topics such as the economy, health care, taxes and election.

We extend the LIAR dataset to the LIAR-PLUS dataset by automatically extracting for each claim the justification that humans have provided in the fact-checking article associated with the claim. Most of the articles end with a summary that has a headline "our ruling" or "summing up". This summary usually has several justification sentences that are related to the statement. We extract all sentences in these summary sections, or the last five sentences in the fact-checking article when no summary exists. We filter out the sentence that has the verdict and related words. These extracted sentences can support or contradict the statement, which is expected to enhance the accuracy of the classification approaches. Excerpt from the LIAR-PLUS dataset is shown in Table 1.

Statement:"Says Rick Scott cut education to pay for even more tax breaks for big, powerful, well-connected corporations."
Speaker: Florida Democratic Party
Context: TV Ad
Label: half-true
Extracted Justification: A TV ad by the Florida Democratic Party says Scott "cut education to pay for even more tax breaks for big, powerful, well-connected corporations." However, the ad exaggerates when it focuses attention on tax breaks for "big, powerful, well-connected corporations." Some such companies benefited, but so did many other types of businesses. And the question of whether the tax cuts and the education cuts had any causal relationship is murkier than the ad lets on.

Table 1: Excerpt from the LIAR-PLUS dataset

3 Methods

The main goal of our paper is to show that modeling the human-provided justification — which can be seen as a summary evidence — improves the assessment of a claim's truth when compared to modeling the claim (and metadata) alone, regardless of the machine learning models (feature based vs. deep learning models). All our models use 4 different conditions: *basic claim/statement*[2] *representation* using just word representations (**S condition**), *enhanced claim/statement representation* that captures additional information shown to be useful such as hedging, sentiment strength and emotion (Rashkin et al., 2017) as well as *metadata information* (**S^{+}M condition**), *basic claim/statement* and the associated *extracted justification* (**SJ condition**) and finally *enhanced claim/statement representation*, *metadata* and *justification* (**S^{+}MJ condition**).

Feature-based Machine Learning. We experiment with both Logistic Regression (LR) and Support Vector Machines (SVM) with linear kernel. For the basic representation of the claim/statement (S condition) we experimented with unigram features, tf-idf weighted unigram features and Glove word embeddings (Pennington et al., 2014). The best representation proved to be unigrams. For the enhanced statement representation (S^{+}) we modeled: sentiment strength using SentiStrength, which measures the negativity and positivity of a statement on a scale of 1-to-5 (Thelwall et al., 2010); emotion using the NRC Emotion Lexicon (EmoLex), which associates each word with eight basic emotions (Mohammad and Turney, 2010), and the Linguistic Inquiry and Word Count (LIWC) lexicon (Pennebaker et al., 2001). In addition, we include metadata information such as the number of claims each speaker makes for every truth-label (history) (Wang, 2017; Long et al., 2017). Finally for representing the justification in the SJ and S^{+}MJ conditions, we just use unigram features.

[2] In the rest of the paper we will refer to the claim as statement.

Cond.	Model	Binary		Six-way	
		valid	test	valid	test
S	LR	0.58	0.61	0.23	0.25
	SVM	0.56	0.59	0.25	0.23
	BiLSTM	0.59	0.60	0.26	0.23
SJ	LR	0.68	0.67	0.37	0.37
	SVM	0.65	0.66	0.34	0.34
	BiLSTM	0.70	0.68	0.34	0.31
	P-BiLSTM	0.69	0.67	0.36	0.35
S^{+}M	LR	0.61	0.61	0.26	0.25
	SVM	0.57	0.60	0.26	0.25
	BiLSTM	0.62	0.62	0.27	0.25
S^{+}MJ	LR	0.69	0.67	0.38	0.37
	SVM	0.66	0.66	0.35	0.35
	BiLSTM	0.71	0.68	0.34	0.32
	P-BiLSTM	0.70	0.70	0.37	0.36

Table 2: Classification Results

Deep Learning Models. We chose to use Bi-Directional Long Short-term Memory (BiLSTM) (Hochreiter and Schmidhuber, 1997) architectures that have been shown to be successful for various related NLP tasks such a textual entailment and argument mining. For the S condition we use just one BiLSTM to model the statement. We use Glove pre-trained word embeddings (Pennington et al., 2014), a 100 dimension embedding layer that is followed by a BiLSTM layer of size 32. The output of the BiLSTM layer is passed to a softmax layer. In the S^{+}M condition, a normalized count vector of those features (described above) is concatenated with the output of the BiLSTM layer to form a merge layer before the softmax. We used a categorical cross_entropy loss function and ADAM optimizer (Kingma and Ba, 2014) and trained the model for 10 epochs. For the SJ and S^{+}MJ conditions we experiment with two architectures: in the first one we just concatenate the justification to the statement and pass it to a single BiLSTM, and in the second one we use a dual/parallel architecture where one BiLSTM reads the statement and another one reads the justification (architecture denoted as P-BiLSTM). The outputs of these BiLSTMs are concatenated and passed to a softmax layer. This latter architecture has been proven to be effective for tasks that model two inputs such as textual entailment (Conneau et al., 2017) or sarcasm detection based on conversation context (Ghosh et al., 2017; Ghosh and Veale, 2017).

Class	class size	S		SJ		
		LR	BiLSTM	LR	BiLSTM	P-BiLSTM
pants-fire	116	0.18	0.19	0.37	0.34	0.37
false	263	0.28	0.34	0.33	0.3	0.33
mostly-false	237	0.21	0.13	0.35	0.31	0.32
half-true	248	0.22	0.28	0.39	0.31	0.37
mostly-true	251	0.23	0.33	0.40	0.39	0.39
true	169	0.22	0.18	0.37	0.42	0.39
total/avg	1284	0.23	0.26	0.37	0.34	0.36

Table 3: F1 Score Per Class on Validation Set

Class	class size	S		SJ		
		LR	BiLSTM	LR	BiLSTM	P-BiLSTM
pants-fire	92	0.12	0.11	0.38	0.33	0.39
false	250	0.31	0.31	0.35	0.32	0.35
mostly-false	214	0.25	0.15	0.35	0.27	0.33
half-true	267	0.24	0.26	0.41	0.27	0.34
mostly-true	249	0.23	0.30	0.35	0.35	0.33
true	211	0.25	0.16	0.37	0.36	0.41
total/avg	1283	0.25	0.23	0.37	0.31	0.35

Table 4: F1 Score Per Class on Test Set

4 Results and Error Analysis

Table 2 shows the results both for the binary and the six-way classification tasks under all 4 conditions (S, SJ, S^{+}M and S^{+}MJ) for our feature-based machine learning models (LR and SVM) and the deep learning models (BiLSTM and P-BiLSTM). For the binary runs we grouped pants on fire, false and mostly false as FALSE and true, mostly true and half true as TRUE. As reference, Wang (2017 best models (text and metadata) obtained 0.277 F1 on validation set and 0.274 F1 on test set in the six-way classification, showing relatively similar results with our equivalent S^{+}M condition.

It is clear from the results shown in Table 2 that including the justification (SJ and S^{+}MJ conditions) improves over the conditions that do not use the justification (S and S^{+}M, respectively) for all models, both in the binary and the six-way classification tasks. For example, for the six-way classification, we see that the BiLSTM model for the SJ condition obtains 0.35 F1 compared to 0.23 F1 in the S condition. LR model has a similar behaviour with 0.37 F1 for the SJ condition compared to 0.25 F1 in S condition. For the S^{+}MJ conditions the best model (LR) shows an F1 of 0.38 compared to 0.26 F1 in the S^{+}M condition (similar results for the deep learning). The dual/parallel BiLSTM architecture provides a small improvement over the single BiLSTM only in the six-way classification.

We also present the per-class results for the six-way classification for the S and SJ conditions. Table 3 shows the results on validation set, while Table 4 on the test set. In the S condition, we

ID	Statement	Justification	label	S	S^+M	SJ	S^+MJ
1	We have the highest tax rate anywhere in the world.	Trump, while lamenting the condition of the middle class, said the U.S. has ”the highest tax rate anywhere in the world.” All sets of data we examined for individual and family taxes prove him wrong. Statutory income tax rates in the U.S. fall around the end of the upper quarter of nations. More exhaustive measures - which compute overall tax burden per person and as a percentage of GDP - show the U.S. either is in the middle of the pack or on the lighter end of taxation compared with other advanced industrialized nations.	false	X	✓	✓	✓
2	“Says Rick Scott cut education to pay for even more tax breaks for big, powerful, well-connected corporations.”	A TV ad by the Florida Democratic Party says Scott ”cut education to pay for even more tax breaks for big, powerful, well-connected corporations.” However, the ad exaggerates when it focuses attention on tax breaks for ”big, powerful, well-connected corporations.” Some such companies benefited, but so did many other types of businesses. And the question of whether the tax cuts and the education cuts had any causal relationship is murkier than the ad lets on.	half-true	X	X	✓	✓
3	Says Donald Trump has given more money to Democratic candidates than Republican candidates.	but public records show that the real estate tycoon has actually contributed around $350,000 more to Republicans at the state and federal level than Democrats. That, however, is a recent development. Fergusons statement contains an element of truth but ignores critical facts.	mostly-false	X	X	✓	✓
4	Says out-of-state abortion clinics have marketed their services to minors in states with parental consent laws.	As Cousins clinic in New York told Yellow Page users in Pennsylvania, ”No state consents.” This is information the clinics wanted patients or potential patients to have, and paid money to help them have it. Whether it was to help persuade them to come in or not, it provided pertinent facts that could help them in their decision-making. It fit the definition of marketing.	true	X	X	X	✓
5	Obamacare provision will allow forced home inspections by government agents.	But the program they pointed to provides grants for voluntary help to at-risk families from trained staff like nurses and social workers. What bloggers describe would be an egregious abuse of the law not whats allowed by it.	pants-fire	X	X	X	✓
6	In the month of January, Canada created more new jobs than we did.	In November 2010, the U.S. economy created 93,000 jobs, compared to 15,200 for Canada. And in December 2010, the U.S. created 121,000 jobs, compared to 22,000 for Canada. ”But on a per capita basis, in recent months U.S. job creation exceeded Canada’s only in October.” January happened to be a month when U.S. job creation was especially low and Canadian job creation was especially high, but it is the most recent month and it reflects the general pattern when you account for population.	true	X	X	X	X
7	There has been $5 trillion in debt added over the last four years.	number is either slightly high or a little low, depending on the type of measurement used, and thats actually for a period short of a full four years. His implication that Obama and the Democrats are to blame has some merit, but it ignores the role Republicans have had.	mostly-true	X	X	X	X

Table 5: Error analysis of Six-way Classification (Logistic Regression)

see a larger degree of variation in performance among classes, with the worst being the pants-on-fire for all models, and for the deep learning model also the mostly-false and true classes. In the SJ condition, we notice a more uniform performance on all classes for all the models. We notice the biggest improvement for the pants-on-fire class for all models, half-true for LR and mostly-false and true for the deep learning models. When comparing the P-BiLSTM and BiLSTM we noticed that the biggest improvement comes from the half-true class and the pants-on-fire class.

Error Analysis In order to further understand the cause of the errors made by the models, we analyzed several examples by looking at the statement, justification and predictions by the logistic regression model when using the S, S^+M, SJ, and S^+MJ conditions (Table 5). Logistic regression was selected since it has the best numbers for the six-way classification task.

The first example in Table 5 was wrongly classified in the S condition, but classified correctly in the S^+M, SJ and S^+MJ conditions. The justification text has a sentence saying “Statutory income tax rates in the U.S. fall around the end of the upper quarter of nations.”, which contradicts the statement and thus is classified correctly when modeling the justification.

The second and third examples in Table 5 were correctly predicted only when the justification was modeled (SJ and S^+MJ conditions). For statement 2, the justification text has a sentence “However, the ad exaggerates...” indicates that the statement has some false and some true information. Therefore, the model predicts the correct label “half-true” when modeling the justification text. Also, the justification for statement 3 was simple enough for the model to predict the gold label “mostly-false”. It has a phrase like “more to Republicans” while the statement had “more to Democratic candidates” which indicates falsehood in the statement as well as discourse markers indicating concessive moves (“but” and “however”).

Sometimes justification features alone were not enough to get the correct prediction without using the *enhanced statement* and metadata features. The justification for statement 4 in Table 5 is complex and no direct connection can be made to the statement. Therefore, the model fails when using SJ and S^+M conditions and only succeed when using all features (i.e., S^+MJ condition). In addition, consider the 5th statement in Table 5 about

Obamacare, it seems that metadata features, which have the history of the speaker, might have helped in predicting its factuality to be “pants on fire”, while it was wrongly classified when modeling only the statement and the justification.

For around half of the instances in validation set, all models had wrong predictions. This is not surprising since the best model had an average F1 score of less than 0.40. The last two example in Table 5 are instances where the model makes mistakes under all 4 conditions. The claim and justification refer to temporal information which is harder to model by the rather simple and shallow approaches we used. Incorporating temporal and numeric information when modeling the claim and justification would be essential for capturing the correct context of a given statement. Another source of error for justification-based conditions was the noise in the extraction of the justification particularly when the “our ruling” and “summing up” headers were not included and we resorted to extract the last 5 sentences from the fact-checking articles. Improving the extraction methods will be helpful to improving the justification-based classification results.

5 Conclusion and Future Work

We presented a study that shows that modeling the human-provided justification form the fact-checking article associated with a claim is important leading to significant improvements when compared to modeling just the claim/statement and metadata for all the machine learning models both in a binary and a six-way classification task. We released LIAR-PLUS, the extended LIAR dataset that contains the automatically extracted justification. We also provided an error analysis and discussion of per-class performance.

Our simple method for extracting the justification from the fact-checking article can lead to slightly noisy text (for example it can contain a repetition of the claim or it can fail to capture the entire evidence). We plan to further refine the justification extraction method so that it contains just the summary evidence. In addition, we plan to develop methods for evidence extraction from the web (similar to the goals of the FEVER shared task (Thorne et al., 2018)) and compare the results of the automatically extracted evidence with the human-provided justifications for fact-checking the claims.

References

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.

Aniruddha Ghosh and Tony Veale. 2017. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 482–491.

Debanjan Ghosh, R. Alexander Fabbri, and Smaranda Muresan. 2017. The role of conversation context for sarcasm detection in online interactions. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 186–196. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Jeffrey Pennington, Richard Socher, and Christopher 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.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240. Association for Computational Linguistics.

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.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology*, 61(12):2544–2558.

J Thorne and A Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*.

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.

William Yang Wang. 2017. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, Vancouver, BC, Canada. ACL.

Xuezhi Wang, Cong Yu, Simon Baumgartner, and Flip Korn. 2018. Relevant document discovery for fact-checking articles. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 525–533. International World Wide Web Conferences Steering Committee.