# It Takes Nine to Smell a Rat: Neural Multi-Task Learning for Check-Worthiness Prediction

**Slavena Vasileva**
Sofia University
slav.vasileva@gmail.com

**Pepa Atanasova**
University of Copenhagen
pepa@di.ku.dk

**Lluís Màrquez**
Amazon Core ML
lluismv@amazon.com

**Alberto Barrón-Cedeño**
Università di Bologna
a.barron@unibo.it

**Preslav Nakov**
Qatar Computing Research Institute, HBKU
pnakov@hbku.edu.qa

## Abstract

We propose a multi-task deep-learning approach for estimating the check-worthiness of claims in political debates. Given a political debate, such as the 2016 US Presidential and Vice-Presidential ones, the task is to predict which statements in the debate should be prioritized for fact-checking. While different fact-checking organizations would naturally make different choices when analyzing the same debate, we show that it pays to learn from multiple sources simultaneously (PolitiFact, FactCheck, ABC, CNN, NPR, NYT, Chicago Tribune, The Guardian, and Washington Post) in a multi-task learning setup, even when a particular source is chosen as a target to imitate. Our evaluation shows state-of-the-art results on a standard dataset for the task of check-worthiness prediction.

## 1 Introduction

Recent years have seen the explosion of fake news, rumors, false claims, distorted facts, half-true statements, and propaganda, which are spreading primarily in social media, but also via standard news broadcasters. This trend became particularly evident during the 2016 US Presidential campaign, which was the turning point that attracted wide public attention to the problem. By then, a number of organizations, e.g., FactCheck[1] and Snopes[2] among many others, launched fact-checking initiatives. Yet, this proved to be a very demanding manual effort, and only a relatively small number of claims could be fact-checked. Thus, it is important to prioritize what to check.

The task of detecting check-worthy claims has been recognized as an important stage in the process of fully automatic fact-checking. According to Vlachos and Riedel (2014) this is a multi-step process that (*i*) extracts statements to be fact-checked, (*ii*) constructs appropriate questions, (*iii*) obtains the answers from relevant sources, and (*iv*) reaches a verdict using these answers. Hassan et al. (2015a) presented a similar vision, and in a follow up work they made check-worthiness an integral part of an end-to-end fact-checking system Hassan et al. (2017).

Here, we approach the problem of mimicking the selection strategy of several renowned fact-checking organizations such as PolitiFact, FactCheck, ABC, CNN, NPR, NYT, Chicago Tribune, The Guardian, and The Washington Post. An important characteristic of this setup is that, perhaps due to editorial policies, fact-checking organizations often select different claims for the same text, with little overlap in their choices (see Tables 1 and 2). Yet, it has been previously shown that it might be beneficial to learn from the union of the selections by multiple fact-checking organizations (Gencheva et al., 2017). Thus, we propose a multi-task deep learning framework, in which we try to predict the choice of each and every fact-checking organization simultaneously. We show that, even when the goal is to mimic the choice of one particular fact-checking organization, it is beneficial to leverage on the choices by multiple such organizations. The evaluation results on a standard dataset show state-of-the-art results.

The remainder of this paper is organized as follows. Section 2 provides an overview of related work. Section 3 describes the used dataset. Section 4 describes our method and features. Section 5 presents the experiments and the evaluation results. Finally, Section 7 concludes and points to some possible directions for future work.

[1] http://www.factcheck.org/
[2] http://www.snopes.com/



*Proceedings of Recent Advances in Natural Language Processing*, pages 1229–1239,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_141

## 2 Related Work

The proliferation of false information has attracted a lot of research interest recently. This includes challenging the truthiness of news (Brill, 2001; Hardalov et al., 2016; Potthast et al., 2018), of news sources (Baly et al., 2018, 2019), and of social media posts (Canini et al., 2011; Castillo et al., 2011; Zubiaga et al., 2016), as well as studying credibility, influence, bias, and propaganda (Ba et al., 2016; Chen et al., 2013; Mihaylov et al., 2015; Kulkarni et al., 2018; Baly et al., 2018; Mihaylov et al., 2018; Barrón-Cedeño et al., 2019; Da San Martino et al., 2019; Zhang et al., 2019).

Research was facilitated by shared tasks such as the SemEval 2017 and 2019 tasks on Rumor Detection (Derczynski et al., 2017; Gorrell et al., 2019), the CLEF 2018 and 2019 CheckThat! labs (Nakov et al., 2018; Elsayed et al., 2019b,a), which featured tasks on automatic identification (Atanasova et al., 2018, 2019) and verification (Barrón-Cedeño et al., 2018; Hasanain et al., 2019) of claims in political debates, the FEVER 2018 and 2019 task on Fact Extraction and VERification (Thorne et al., 2018), and the SemEval 2019 task on Fact-Checking in Community Question Answering Forums (Mihaylova et al., 2019), among others.

The interested reader can learn more about "fake news" from the overview by Shu et al. (2017), which adopted a data mining perspective and focused on social media. Another recent survey (Thorne and Vlachos, 2018) took a fact-checking perspective on "fake news" and related problems. Yet another survey was performed by Li et al. (2016), and it covered truth discovery in general. Moreover, there were two recent articles in *Science*: Lazer et al. (2018) offered a general overview and discussion on the science of "fake news", while Vosoughi et al. (2018) focused on the proliferation of true and false news online.

The first work to target check-worthiness estimation, i.e., predicting which sentences in a given input text should be prioritized for fact-checking, was the ClaimBuster system (Hassan et al., 2015b). It is trained on data that was manually annotated by students, professors, and journalists, where each sentence was marked as *non-factual*, *unimportant factual*, or *check-worthy factual*. The system used an SVM classifier and features such as sentiment, TF.IDF representations, part-of-speech tags, and named entities.

In our previous work (Gencheva et al., 2017; Jaradat et al., 2018), we used debates from the 2016 US Presidential Campaign and fact-checking reports by professional journalists; we use this same dataset here. Beside most of the features borrowed from ClaimBuster, our model paid special attention to the context of each sentence. This includes whether it is part of a long intervention by one of the debate participants and its position within such an intervention. We predicted both (*i*) whether any of the fact-checking organizations would select the target sentence, and also (*ii*) whether a specific fact-checking organization would select it. There was also a lab on fact-checking at CLEF 2018 and 2019 (Atanasova et al., 2018, 2019), which was partially based on a variant of this data, but it focused on one fact-checking organization, unlike our multi-source setup here.

Patwari et al. (2017) also focused on the 2016 US Election campaign. Their setup asks to predict whether any of the fact-checking organizations would select the target sentence. They used a boosting-like model that takes SVMs focusing on different clusters of the dataset and the final outcome is considered as that coming from the most confident classifier. The features considered range from LDA topic-modeling to part-of-speech (POS) tuples and bag-of-words representations.

Other claim monitoring tools include FactWatcher (Hassan et al., 2014) and DisputeFinder (Ennals et al., 2010b). FactWatcher classifies claims as situational facts, one-of-the-few, or prominent streaks. It checks whether a new text triggers some of the three types of claims, treating the sentences in the text as sequential data. DisputeFinder mines the Web for already-verified claims. Both maintain a growing database of facts and known claims.

Beyond the document context, it has been proposed to mine check-worthy claims on the Web. For example, Ennals et al. (2010a) searched for linguistic cues of disagreement between the author of a statement and what is believed, e.g., "*falsely claimed that X*". The claims matching the patterns would then go through a classifier. This procedure can be used to acquire a dataset of disputed claims.

Given a set of disputed claims, Ennals et al. (2010b) looked for new claims on the Web that entail the ones that have already been collected. Thus, the task can be reduced to recognizing textual entailment (Dagan et al., 2009).

de Marneffe et al. (2008) also looked for contradictions in text. They tried to classify the contradictions that can be found in a piece of text in two categories —those occurring via antonymy, negation, and date/number mismatch, and those arising from different world knowledge and lexical contrasts. The features that are selected for the task of contradiction detection include polarity, numbers, dates and time, antonymy, factivity, modality, structural, and relational features.

Finally, Le et al. (2016) used deep learning. They argued that the top terms in claim vs. non-claim sentences are highly overlapping in content, which is a problem for bag-of-words approaches. Thus, they used a Convolutional Neural Network, where each word is represented by its embedding and each named entity is replaced by its tag, e.g., *person*, *organization*, *location*.

Unlike the above work, we mimic the selection strategy of *one* specific fact-checking organization by learning to jointly predict the selection choices by *multiple* such organizations.

## 3 Data

In our experiments, we used the CW-USPD-2016 dataset from our previous work (Gencheva et al., 2017), which can be found on GitHub.[3] It is derived from transcripts of the 2016 US Presidential campaign, and includes one Vice-Presidential and three Presidential debates, all of which were fact-checked by the following nine reputable fact-checking organizations: PolitiFact, FactCheck, ABC, CNN, NPR, NYT, Chicago Tribune, The Guardian, and The Washington Post.

Overall, there are four debates with a total of 5,415 sentences. A sentence is considered check-worthy with respect to a source if that source has chosen to fact-check it. Overall, a total of 880 sentences were fact-checked by at least one source, 191 were selected by two or more sources, 100 by three or more, and only one sentence was chosen by all nine sources, as Table 1 shows. Table 2 shows an example: interventions by Hillary Clinton and Donald Trump from the first US presidential debate. This reflects the disparities in check-worthiness selection criteria. More details about the dataset can be found in (Gencheva et al., 2017).

[3] http://github.com/pgencheva/claim-rank

| Selected by # Sources | Number of Sentences | Cumulative Sum |
|---|---|---|
| 9 | 1 | 1 |
| 8 | 6 | 7 |
| 7 | 5 | 12 |
| 6 | 19 | 31 |
| 5 | 26 | 57 |
| 4 | 40 | 97 |
| 3 | 100 | 197 |
| 2 | 191 | 388 |
| **1** | **492** | **880** |

Table 1: Agreement between the fact-checkers: sentences selected by 1, 2, . . ., 9 of them.

## 4 Our Multi-Task Learning Model

We approach the task of check-worthiness prediction as a multi-source learning problem, using different sources of annotation over the same training dataset. Thus, we can learn to mimic the selection strategy of each of the individual sources.

Figure 1 shows the architecture of our neural multi-task learning model which, given an input sentence in the context of a political debate, predicts whether each of the nine individual sources (tasks) would have selected it, and whether at least one of them would, which is the special *task ANY*.

The input to our neural network consists of various domain-specific features that have been previously shown to work well for the task of check-worthiness prediction. In particular, from (Hassan et al., 2015b), we adopt TF.IDF-weighted bag of words, part-of-speech tags, the presence of named entities, sentiment scores, and sentence length (in number of tokens). Moreover, from (Gencheva et al., 2017), we further adopt *lexicon features*, e.g., for bias (Recasens et al., 2013), for sentiment (Liu et al., 2005), for assertiveness (Hooper, 1974), and for subjectivity; *structural features*, e.g., for location of the sentence within the debate/intervention; LDA topics (Blei et al., 2003); word embeddings, pre-trained on Google News (Mikolov et al., 2013); and discourse relations with respect to the neighboring sentences (Joty et al., 2015). See (Hassan et al., 2015b; Gencheva et al., 2017) for more details about each of these feature types.

After the input layer, comes a hidden layer that is shared between all tasks. It is followed by ten parallel task-specific hidden layers. During training, in the process of backpropagation, each task modifies the weights of its own task-specific layer and also of the shared layer.

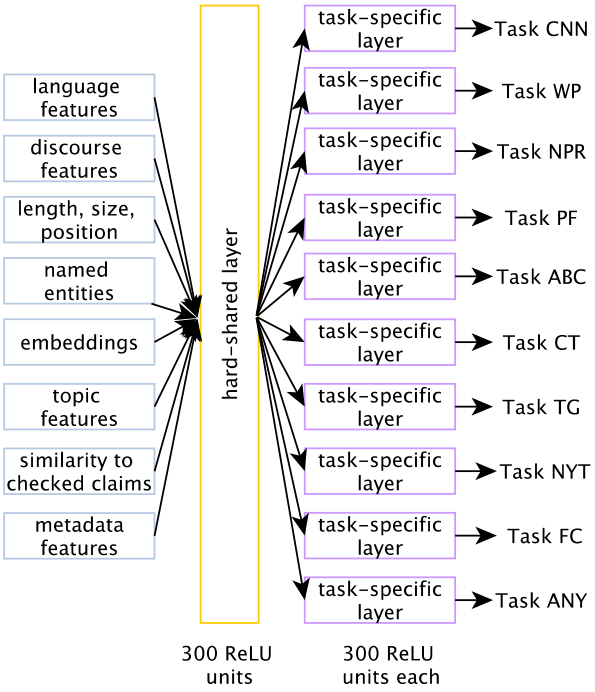


Figure 1: The architecture of our neural multi-task learning model, predicting whether each of the nine individual fact-checking organizations (tasks) would consider this sentence check-worthy and one cumulative source: *task ANY*.

Finally, each task-specific layer is followed by an output layer: a single sigmoid unit that provides the prediction of whether the utterance was fact-checked by the corresponding source. Eventually, we make use of the probability of the prediction to prioritize claims for fact-checking. This kind of neural network architecture for multi-task learning is known in the literature as *hard parameter sharing* (Caruana, 1993), and it can greatly reduce the risk of overfitting.

## 5 Experiments and Evaluation

As the CW-USPD-2016 corpus contains four debates, we perform 4-fold cross-validation, where each time we leave one debate out for testing, and we train on the remaining three debates. Moreover, in order to stabilize the results, we repeat each experiment three times with different random seeds and we report the average over these three reruns of the system.[4]

[4]Having multiple reruns is a standard procedure to stabilize an optimization algorithm that is sensitive to the random seed, e.g., this strategy has been argued for when using MERT for tuning hyper-parameters in Statistical Machine Translation (Foster and Kuhn, 2009).

In our neural model, we used ReLU units and a shared layer of size 300. For training, we used Stochastic Gradient Descent with Nesterov momentum,[5] iterating for 100 epochs.

Recall that our main objective is to prioritize the claims that should be selected for manual fact-checking, which is best achieved by proposing a ranked list of claims. Thus, we have a ranking task, for which we use suitable information retrieval evaluation measures. In particular, we adopt Mean Average Precision (MAP) as our primary evaluation measure. We further report R-Precision, or R-Pr, and precision at $k$, or P@$k$,[6] for $k = \{5, 10, 20, 50\}$. Note that 50 is the approximate number of claims checked by most of the sources for each debate (the exception being PolitiFact, with up to 99 checked claims).

Table 3 presents the evaluation results comparing three models. The first one is a single-task model *singleton* where a separate neural network is trained for each source. The other two are multi-task learning models: *multi* predicts labels for each of the nine tasks, one for each fact-checker, and *multi+any* predicts labels for each of the nine tasks (one for each fact-checker), and also for *task ANY* (as shown in Figure 1). We further compare to the online version of ClaimBuster (Hassan et al., 2015b) and to the *singleton* results reported in (Gencheva et al., 2017) (on the same dataset, with the same cross-validation).[7]

We can see in Table 3 that our *singleton* is comparable and even slightly better than the *singleton* model in (Gencheva et al., 2017), and both outperform the online version of ClaimBuster (Hassan et al., 2015a). We further see that limiting our singleton system to ClaimBuster's features yields a sizable drop in performance. Moreover, for most sources, multi-task learning improves over the singleton models. The results of the multi-task variations that improve over the single baseline are boldfaced. The improvements are consistent across the evaluation measures, but they vary largely depending on the fact-checking source and the evaluation measure.

[5]Using Adam optimizer was faster, converging after only 30 epochs, but it yielded slightly worse results.

[6]See (Buckley and Voorhees, 2000) for a discussion on these evaluation measures.

[7]Note that we could not compare to (Patwari et al., 2017) directly as they used a different dataset. However, they use a small set of basic features that overlap with those of ClaimBuster (Hassan et al., 2015b) to a large extent, and thus we expect that they would perform similarly to ClaimBuster.

| Speaker | Total | CT | ABC | CNN | WP | NPR | PF | TG | NYT | FC | Text |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Clinton | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | So we're now on the precipice of having a potentially much better economy, but the last thing we need to do is to go back to the policies that failed us in the first place. |
| Clinton | 6 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | Independent experts have looked at what I've proposed and looked at what Donald's proposed, and basically they've said this, that if his tax plan, which would blow up the debt by over $5 trillion and would in some instances disadvantage middle-class families compared to the wealthy, were to go into effect, we would lose 3.5 million jobs and maybe have another recession. |
| Clinton | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | They've looked at my plans and they've said, OK, if we can do this, and I intend to get it done, we will have 10 million more new jobs, because we will be making investments where we can grow the economy. |
| Clinton | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Take clean energy. |
| Clinton | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Some country is going to be the clean- energy superpower of the 21st century. |
| Clinton | 6 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | Donald thinks that climate change is a hoax perpetrated by the Chinese. |
| Clinton | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | I think it's real. |
| Trump | 5 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | I did not. |

Table 2: Excerpt from the transcript of the first US 2016 Presidential Debate, annotated by nine sources: Chicago Tribune, ABC News, CNN, Washington Post, NPR, PolitiFact, The Guardian, The New York Times and Factcheck.org. Whether the media fact-checked the claim or not is indicated by a 1 or 0, respectively. The blue sentences are considered as positive in the *any* setting.

One notable exception is NYT, for which the single-task learning shows the highest scores. We hypothesize that the network has found some distinctive features of NYT, which make it easy to predict. These relations are blurred when we try to optimize for multiple tasks at once. However, it is important to state that removing NYT from the learning targets worsens the results for the other sources, i.e. it carries some important relations that are worth modeling.

Table 4 presents the same results but averaged over the nine sources. The first section in Table 4 shows the results for the online version of ClaimBuster (Hassan et al., 2015b), and for the *singleton* and the *task ANY* results in (Gencheva et al., 2017). We can see that our *singleton* model is comparable to the *singleton* and *any* models in (Gencheva et al., 2017), and our multi-task learning models consistently improve over them for all evaluation measures in all but one case.

It is common in neural networks to try to implicitly learn the representations based on word embeddings. We include this as a baseline in the second section in Table 4. The performance of the model that only uses embeddings is in general poor, which suggests that complex feature modeling is necessary for this task; including features that go beyond the current-sentence level. Further feature analysis is included in Table 6.

The third section of Table 4 presents the results for the models of this paper. Again, we can see that multi-task learning yields sizeable improvement over the single-task learning baseline for all evaluation measures.

Another conclusion that can be drawn from this table is that including the task *task ANY* (i.e., whether any of the nine media would select a target) does not help to improve the multi-task model. This is probably due to the fact that this information is already contained in the multi-task model with nine sources.

The last section in Table 4 presents two additional variants of the model: the single-task learning *any* system —which trains on the union of the selected sentences by all nine fact-checkers to predict the target fact-checker only—, and the system *singleton+any* that predicts labels for two tasks: (*i*) for the target fact-checker, and (*ii*) for *task ANY*. We can see that *any* performs comparably to the *singleton* baseline, thus being clearly inferior than the multi-task learning variants. Finally, *singleton+any* is also better than the single-task learning variants, but it falls short compared to the other multi-task learning variants. Including output units for all nine individual media seems crucial for getting advantage of the multi-task learning, i.e., considering only an extra output prediction node for the *task ANY* problem is not enough.

| Model | MAP | R-Pr | P@5 | P@10 | P@20 | P@50 |
|---|---|---|---|---|---|---|
| **ABC** | | | | | | |
| singleton CB | .057 | .061 | .050 | .038 | .056 | .050 |
| CB online | .065 | .066 | .150 | .125 | .088 | .080 |
| singletonG | .059 | .068 | .050 | .050 | .100 | .060 |
| *singleton* | .097 | .112 | .250 | .175 | .162 | .100 |
| ***multi*** | **.119** | **.157** | **.333** | **.225** | **.217** | **.122** |
| ***multi+any*** | **.118** | **.160** | **.300** | **.233** | **.229** | **.132** |
| **The Washington Post (WP)** | | | | | | |
| singleton CB | .051 | .053 | .050 | .033 | .046 | .048 |
| CB online | .048 | .056 | .050 | .075 | .050 | .045 |
| singletonG | .102 | .098 | .200 | .175 | .113 | .080 |
| *singleton* | .106 | .110 | .150 | .100 | .112 | .110 |
| ***multi*** | **.127** | **.127** | **.350** | **.233** | **.162** | **.123** |
| ***multi+any*** | **.130** | **.129** | **.350** | **.250** | **.171** | ***.110*** |
| **CNN** | | | | | | |
| singleton CB | .055 | .058 | .063 | .038 | .050 | .053 |
| CB online | .082 | .096 | .150 | .125 | .088 | .085 |
| singletonG | .079 | .076 | .100 | .100 | .100 | .090 |
| *singleton* | .087 | .091 | .250 | .150 | .121 | .090 |
| ***multi*** | **.113** | **.132** | ***.250*** | **.208** | **.183** | **.140** |
| ***multi+any*** | **.109** | **.126** | .167 | **.200** | **.167** | **.128** |
| **FactCheck (FC)** | | | | | | |
| singleton CB | .068 | .072 | .108 | .071 | .077 | .070 |
| CB online | .081 | .213 | .150 | .125 | .100 | .115 |
| singletonG | .081 | .098 | .050 | .125 | .088 | .085 |
| *singleton* | .084 | .114 | .117 | .125 | .088 | .100 |
| ***multi*** | **.105** | **.136** | **.250** | **.175** | **.146** | **.118** |
| ***multi+any*** | **.117** | .110 | **.333** | **.242** | **.196** | **.107** |
| **PolitiFact** | | | | | | |
| singleton CB | .137 | .143 | .250 | .200 | .188 | .185 |
| CB online | .154 | .213 | .200 | .300 | .238 | .210 |
| singletonG | .218 | .274 | .450 | .325 | .300 | .270 |
| *singleton* | .201 | .278 | .250 | .250 | .262 | .262 |
| ***multi*** | **.209** | .258 | **.400** | **.367** | **.317** | **.270** |
| ***multi+any*** | **.210** | .252 | **.500** | **.350** | **.333** | **.272** |

| Model | MAP | R-Pr | P@5 | P@10 | P@20 | P@50 |
|---|---|---|---|---|---|---|
| **NPR** | | | | | | |
| singleton CB | .079 | .085 | .136 | .089 | .096 | .087 |
| CB online | .144 | .186 | .200 | .225 | .225 | .180 |
| singletonG | .193 | .216 | .550 | .475 | .350 | .255 |
| *singleton* | .175 | .195 | .250 | .250 | .283 | .228 |
| ***multi*** | **.186** | **.210** | **.333** | **.342** | **.300** | **.245** |
| ***multi+any*** | **.180** | **.207** | **.333** | **.283** | .250 | .227 |
| **The Guardian (TG)** | | | | | | |
| singleton CB | .066 | .075 | .110 | .070 | .070 | .066 |
| CB online | .084 | .128 | .100 | .100 | .125 | .140 |
| singletonG | .121 | .156 | .250 | .225 | .200 | .155 |
| *singleton* | .127 | .174 | .200 | .150 | .196 | .178 |
| ***multi*** | **.133** | **.199** | .183 | **.175** | **.192** | **.193** |
| ***multi+any*** | **.130** | .159 | **.217** | **.175** | **.200** | .167 |
| **Chicago Tribune (CT)** | | | | | | |
| singleton CB | .058 | .063 | .050 | .050 | .050 | .065 |
| CB online | .053 | .032 | .050 | .050 | .038 | .065 |
| singletonG | .087 | .118 | .150 | .150 | .175 | .105 |
| *singleton* | .079 | .110 | .100 | .100 | .125 | .075 |
| ***multi*** | **.081** | .090 | ***.100*** | **.133** | .104 | **.082** |
| ***multi+any*** | **.087** | .087 | **.133** | ***.100*** | .108 | **.093** |
| **The New York Times (NYT)** | | | | | | |
| singleton CB | .080 | .084 | .138 | .094 | .100 | .088 |
| CB online | .103 | .250 | .250 | .163 | .135 | .135 |
| singletonG | .136 | .178 | .250 | .225 | .188 | .135 |
| *singleton* | .187 | .221 | .350 | .325 | .238 | .192 |
| ***multi*** | .150 | .213 | .233 | .200 | .196 | .180 |
| ***multi+any*** | .147 | .197 | .200 | .167 | .158 | .162 |

| | |
|---|---|
| singleton CB | Singleton only w/ClaimBuster features |
| CB online | Online version of ClaimBuster |
| singletonG | Singleton from (Gencheva et al., 2017) |
| *singleton* | Trained on the target medium only |
| ***multi*** | Multi-task for nine sources |
| ***multi+any*** | Multi-task for nine sources+any |

Table 3: Evaluation results for each of the nine fact-checking sources as a target to mimic. Shown are the results for single-source baselines vs. for multi-task learning with nine and with ten classes. The improvements over the singleton baseline are marked in bold. We further compare to *singleton* that is limited to ClaimBuster's features, to the online version of ClaimBuster (Hassan et al., 2015b), and to *singletonG* results in (Gencheva et al., 2017). The improvements over the latter are underlined.

| Model | MAP | R-Pr | P@5 | P@10 | P@20 | P@50 |
|---|---|---|---|---|---|---|
| CB online | .090 | .138 | .144 | .143 | .121 | .117 |
| singletonG | .120 | .142 | .228 | .206 | .179 | .137 |
| *anyG* | .128 | .225 | .194 | .186 | .178 | .153 |
| *singleton (embed.)* | .058 | .065 | .055 | .055 | .068 | .072 |
| singleton CB | .072 | .077 | .106 | .076 | .081 | .079 |
| *singleton* | .127 | .156 | .213 | .181 | .176 | .148 |
| ***multi*** | **.136** | **.169** | **.270** | **.229** | **.202** | **.164** |
| ***multi+any*** | **.136** | **.159** | **.281** | **.222** | **.201** | **.155** |
| *any* | .125 | .153 | .204 | .197 | .175 | .153 |
| *singleton+any* | **.130** | .153 | **.237** | **.220** | **.184** | ***.148*** |

Table 4: Evaluation results averaged over nine fact-checking organizations (see Table 3 for the unrolled results). We compare multi-task learning to three *singleton* baselines; the improvements are shown in bold. The first section compares to the online version of ClaimBuster (Hassan et al., 2015b), as well as to *singleton* and to *task ANY* results in (Gencheva et al., 2017). The improvements over the latter are underlined. The last section shows the results for two more baselines: *any* and *singleton+any* .

| N | Type | Tgt | # | Sentence |
|---|---|---|---|---|
| 1 | TP | CT | 8 | **Trump** ▸ It's gone, $6 billion. |
| 2 | TP | WP | 8 | **Trump** ▸ I was against – I was against the war in Iraq. |
| 3 | TP | TG | 6 | **Trump** ▸ You ran the State Department, $6 billion was either stolen. |
| 4 | TP | NYT | 6 | **Pence** ▸ Less than 10 cents on the dollar of the Clinton Foundation has gone to charitable causes. |
| 5 | FP | CT | 4 | **Trump** ▸ Wrong. |
| 6 | FP | CT | 3 | **Trump** ▸ In Chicago, they've had thousands of shootings, thousands since January 1st. |
| 7 | FP | CNN | 0 | **Clinton** ▸ Donald has said he's in favor of defending Planned Parenthood. |
| 8 | FP | WP | 0 | **Trump** ▸ I never met Putin. |
| 9 | FN | FC | 6 | **Clinton** ▸ Donald thinks that climate change is a hoax perpetrated by the Chinese. |
| 10 | FN | NYT | 4 | **Pence** ▸ And Iraq has been overrun by ISIS, because Hillary Clinton failed to renegotiate... |
| 11 | FN | NPR | 1 | **Trump** ▸ China should go into North Korea. |
| 12 | FN | NPR | 1 | **Trump** ▸ We have no growth in this country. |

Table 5: Sentences with prediction type (for the *multi* model, with respect to the target medium), the target medium, and total number of media that selected this sentence (#).

## 6 Discussion

In this section, we provide deeper insight into the peculiar characteristics of the multi-task model.

**Error Analysis** First, we perform comparative error analysis, showing both examples of improvement of the proposed *multi* model with respect to the *singleton* as well as some cases where the former fails. The results are shown in Table 5. The first four rows are true positive claims, which were misclassified by the *singleton* model, but were correctly classified by the *multi-task* one. As we can see, the claims were selected for fact-checking by many organizations: between six and eight. This reflects that these instances were certainly check-worthy and the multi-task model correctly spotted them. The observation holds for a prevailing number of all of the new true positives. This is a natural consequence of our neural architecture, where all sources share a hidden layer and tend to learn from the selection criteria of the other sources as well.

Two types of false positive errors occur in rows 5–8. Rows 5 and 6 are predicted by multiple sources that reinforce one another for the wrong guess. We can attribute this to the specifics of the multi-task architecture. On the one hand, the shared layer helps a medium to learn from the selection process of other media. On the other hand, it begins to make more mistakes on claims selected by more media.

On the contrary, rows 7 and 8 show claims that are not check-worthy for any source, but exhibit features such as named entities and negations that typically suggest that the claim might be check-worthy. Finally, rows 9–12 are false negative instances. We have two claims that were fact-checked by several media and two selected by one medium only. The first group indicates that some tasks might try to learn their own features, while the second group shows a possible down side of the multi-task model.

**Feature Importance** Next, we conduct feature ablation experiments to determine which of the feature groups are most important for the final multi-task model. For this purpose, we remove one feature group at a time from the *multi* model.

Table 6 shows that without the Embedding features the performance of the model drops significantly. They were also the best features in the *singletonG* model of Gencheva et al. (2017). Metadata features are the second most important for the model. An interesting observation is that some of the best-preforming features from *singletonG* are the least contributing to the multi-task model. Such features are *Sim. to prev.* (similarity to previously fact-checked claims), and the linguistic features.

| Feature | MAP | R-Pr | P@5 | P@10 | P@20 | P@50 |
|---|---|---|---|---|---|---|
| Embeddings | .102 | .133 | .250 | .231 | .188 | .129 |
| Metadata | .120 | .147 | .278 | .217 | .175 | .139 |
| Sentiment | .122 | .146 | .233 | .203 | .164 | .140 |
| Topics | .123 | .147 | .244 | .211 | .172 | .142 |
| Discourse | .123 | .140 | .261 | .217 | .175 | .141 |
| NER | .125 | .149 | .244 | .217 | .178 | .140 |
| Segment size | .125 | .149 | .256 | .211 | .172 | .139 |
| Position | .125 | .143 | .261 | .219 | .193 | .138 |
| Linguistic | .126 | .150 | .250 | .208 | .190 | .151 |
| Contradiction | .126 | .149 | .250 | .203 | .174 | .142 |
| Lengths | .127 | .144 | .272 | .233 | .175 | .147 |
| Sim. to prev. | .127 | .151 | .222 | .214 | .178 | .148 |

Table 6: Ablation experiments: removing a feature group from the *multi* model, using all nine tasks.

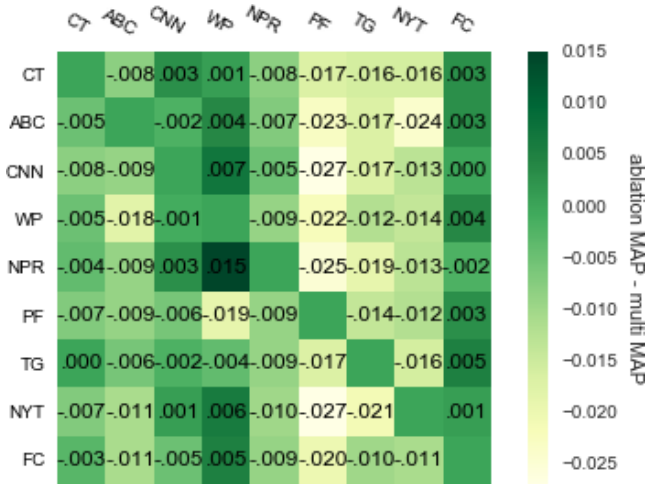


Figure 2: Ablation experiment with the *multi* model. Each row is an experiment removing one target. Each column is the MAP *difference* with respect to the *multi* model for the corresponding target.

**Source Ablation** Figure 2 shows ablation results with the *multi* model. A cell at row $r$ and column $c$ shows the performance difference for target $c$ when excluding the target $r$ at training time. For example, in the first row we run the *multi* model neglecting CT in the set of targets. Negative values indicate that removing target $r$ worsens the MAP of target $c$. Conversely, positive values indicate that removing target $r$ improves MAP for target $c$. We can observe that the MAP of ABC has dropped by .008, meaning that ABC finds beneficial information from sharing a layer with the CT target. On the contrary, the target FC improves after removing CT, pointing out the presence of conflicts in the learning phase of the shared layer. The largest decrease in MAP is observed in PF after removing CNN, NYT, and NPR. On the other hand, the most significant increase in MAP is in WP after removing NPR and CNN.

## 7 Conclusion and Future Work

We have presented a multi-task learning approach for estimating the check-worthiness of claims in political debates, and we have further demonstrated its effectiveness experimentally, pushing the state of the art.

In future work, we plan to experiment with more debates. We further plan to go beyond debates, i.e., to general news articles. Moreover, we would like to apply our approach to other languages for which multiple check-worthiness annotations of the same dataset are available.

We plan to try information sources such as the Web (Popat et al., 2017), as well as tweets and temporal information (Ma et al., 2016). We also want to explore other multi-task learning options, e.g., as described in (Ruder, 2017).

It would be interesting to investigate the reasons why the NYT source does not benefit from the multi-task architecture. In order to adapt to this situation with a single model, we plan to experiment with a network with *soft parameter sharing*, e.g., as in (Duong et al., 2015). For example, we could create a chain of layers that back-propagate to the input using only single task targets and then add an auxiliary layer that is shared between the tasks on the side. In this way, the model would be able to turn off the multi-task learning completely for some of the sources. However, training such kind of model might require significantly more training data; semi-supervised training might be a possible solution.

## Acknowledgments

We would like to thank the anonymous reviewer, whose constructive feedback has helped us improve the quality of this paper.

This work is part of the Tanbih project,[8] which aims to limit the effect of "fake news", propaganda and media bias by making users aware of what they are reading. The project is developed in collaboration between the Qatar Computing Research Institute (QCRI), HBKU and the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL).

[8] http://tanbih.qcri.org/

## References


Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims, Task 1: Check-worthiness. In *CLEF 2018 Working Notes*. CEUR-WS.org, Avignon, France.

Pepa Atanasova, Preslav Nakov, Georgi Karadzhov, Mitra Mohtarami, and Giovanni Da San Martino. 2019. Overview of the CLEF-2019 CheckThat! Lab on Automatic Identification and Verification of Claims. Task 1: Check-Worthiness. In *CLEF 2019 Working Notes*. CEUR-WS.org, Lugano, Switzerland.

Mouhamadou Lamine Ba, Laure Berti-Equille, Kushal Shah, and Hossam M. Hammady. 2016. VERA: A platform for veracity estimation over web data. In *Proceedings of the 25th International Conference Companion on World Wide Web*. Montréal, Québec, Canada, WWW '16, pages 159–162.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, EMNLP '18, pages 3528–3539.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Minneapolis, MN, USA, NAACL-HLT '19, pages 2109–2116.

Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management* 56(5):1849 – 1864.

Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims, Task 2: Factuality. In *CLEF 2018 Working Notes*. CEUR-WS.org, Avignon, France.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3(Jan):993–1022.

Ann M Brill. 2001. Online journalists embrace new marketing function. *Newspaper Research Journal* 22(2):28.

Chris Buckley and Ellen M. Voorhees. 2000. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Athens, Greece, SIGIR '00, pages 33–40.

Kevin R. Canini, Bongwon Suh, and Peter L. Pirolli. 2011. Finding credible information sources in social networks based on content and social structure. In *Proceedings of the IEEE International Conference on Privacy, Security, Risk, and Trust, and the IEEE International Conference on Social Computing*. Boston, MA, USA, SocialCom/PASSAT '11, pages 1–8.

Richard Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the International Conference on Machine Learning*. Amherst, MA, USA, ICML '13, pages 41–48.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the International Conference on World Wide Web*. Hyderabad, India, WWW '11, pages 675–684.

Cheng Chen, Kui Wu, Venkatesh Srinivasan, and Xudong Zhang. 2013. Battling the Internet Water Army: detection of hidden paid posters. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Niagara, Ontario, Canada, ASONAM '13, pages 116–120.

Giovanni Da San Martino, Seunghak Yu, Alberto Barron-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Hong Kong, China, EMNLP '19.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering* 15(4):i–xvii.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Columbus, OH, USA, ACL '08, pages 1039–1047.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 60–67.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network

parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, ACL-IJCNLP '15, pages 845–850.

Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Pepa Atanasova, and Giovanni Da San Martino. 2019a. CheckThat! at CLEF 2019: Automatic identification and verification of claims. In *Proceedings of the 41st European Conference on Information Retrieval*. Cologne, Germany, ECIR '19, pages 309–315.

Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019b. Overview of the CLEF-2019 CheckThat!: Automatic identification and verification of claims. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer, Lugano, Switzerland.

Rob Ennals, Dan Byler, John Mark Agosta, and Barbara Rosario. 2010a. What is disputed on the web? In *Proceedings of the 4th Workshop on Information Credibility*. New York, NY, USA, WICOW '10, pages 67–74.

Rob Ennals, Beth Trushkowsky, and John Mark Agosta. 2010b. Highlighting disputed claims on the web. In *Proceedings of the International Conference on World Wide Web*. New York, NY, USA, WWW '10, pages 341–350.

George Foster and Roland Kuhn. 2009. Stabilizing minimum error rate training. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece, StatMT '09, pages 242–249.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Varna, Bulgaria, RANLP '17, pages 267–276.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, MN, USA, SemEval '19, pages 845–854.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2016. In search of credible news. In *Proceedings of the 17th International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. Varna, Bulgaria, AIMSA '16, pages 172–180.

Maram Hasanain, Reem Suwaileh, Tamer Elsayed, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Overview of the CLEF-2019 CheckThat! Lab on Automatic Identification and Verification of Claims. Task 2: Evidence and Factuality. In *CLEF 2019 Working Notes*. CEUR-WS.org, Lugano, Switzerland.

Naeemul Hassan, Bill Adair, James T. Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015a. The quest to automate fact-checking. In *Proceedings of the Computation+Journalism Symposium*.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015b. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*. Melbourne, Australia, CIKM '15, pages 1835–1838.

Naeemul Hassan, Afroza Sultana, You Wu, Gensheng Zhang, Chengkai Li, Jun Yang, and Cong Yu. 2014. Data in, fact out: Automated monitoring of facts by FactWatcher. *PVLDB* 7:1557–1560.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. ClaimBuster: The first-ever end-to-end fact-checking system. *Proc. VLDB Endow.* 10(12):1945–1948.

Joan B. Hooper. 1974. *On Assertive Predicates*. Indiana University Linguistics Club.

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 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics*. New Orleans, LA, USA, NAACL-HLT '18, pages 26–30.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Comput. Linguist.* 41(3):385–435.

Vivek Kulkarni, Junting Ye, Steve Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, EMNLP '18, pages 3518–3527.

David M.J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359(6380):1094–1096.

Dieu-Thu Le, Ngoc Thang Vu, and Andre Blessing. 2016. Towards a text analysis system for political debates. *LaTeCH 2016* page 134.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A survey on truth discovery. *SIGKDD Explor. Newsl.* 17(2):1–16.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion Observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*. New York, NY, USA, WWW '05, pages 342–351.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. New York, New York, USA, IJCAI '16, pages 3818–3824.

Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. 2015. Finding opinion manipulation trolls in news community forums. In *Proceedings of the Conference on Computational Natural Language Learning*. Beijing, China, CoNLL '15, pages 310–314.

Todor Mihaylov, Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Georgi Georgiev, and Ivan Koychev. 2018. The dark side of news community forums: Opinion manipulation trolls. *Internet Research* 28(5):1292–1312.

Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. SemEval-2019 task 8: Fact checking in community question answering forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, MN, USA, SemEval '19, pages 860–869.

Tomas Mikolov, Wen-Tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. Atlanta, GA, USA, NAACL-HLT '13, pages 746–751.

Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. In *Proceedings of CLEF*. Avignon, France, pages 372–387.

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*. Singapore, CIKM '17, pages 2259–2262.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*. Perth, Australia, WWW '17 Companion, pages 1003–1012.

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*. Melbourne, Australia, ACL '18, pages 231–240.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, ACL '13, pages 1650–1659.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *CoRR* abs/1706.05098.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.* 19(1):22–36.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the International Conference on Computational Linguistics*. Santa Fe, NM, USA, COLING '18, pages 3346–3359.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. New Orleans, LA, USA, NAACL-HLT '18, pages 809–819.

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*. Baltimore, MD, USA, pages 18–22.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359(6380):1146–1151.

Yifan Zhang, Giovanni Da San Martino, Alberto Barrn-Cedeo, Salvatore Romeo, Jisun An, Haewoon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhov, Ramy Baly, Kareem Darwish, and Preslav Nakov James Glass. 2019. Tanbih: Get to know what you are reading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Hong Kong, China, EMNLP '19.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE* 11(3):1–29.