# Automatic Fact-Checking Using Context and Discourse Information

PEPA ATANASOVA, Sofia University "St. Kliment Ohridski," Bulgaria
PRESLAV NAKOV, Qatar Computing Research Institute, HBKU, Qatar
LLUÍS MÀRQUEZ, Amazon, Spain
ALBERTO BARRÓN-CEDEÑO, Qatar Computing Research Institute, HBKU, Qatar
GEORGI KARADZHOV and TSVETOMILA MIHAYLOVA, Sofia University "St. Kliment Ohridski," Bulgaria
MITRA MOHTARAMI and JAMES GLASS, Massachusetts Institute of Technology, USA

We study the problem of automatic fact-checking, paying special attention to the impact of contextual and discourse information. We address two related tasks: (*i*) detecting check-worthy claims and (*ii*) fact-checking claims. We develop supervised systems based on neural networks, kernel-based support vector machines, and combinations thereof, which make use of rich input representations in terms of discourse cues and contextual features. For the check-worthiness estimation task, we focus on political debates, and we model the target claim in the context of the full intervention of a participant and the previous and following turns in the debate, taking into account contextual meta information. For the fact-checking task, we focus on answer verification in a community forum, and we model the veracity of the answer with respect to the entire question–answer thread in which it occurs as well as with respect to other related posts from the entire forum. We develop annotated datasets for both tasks and we run extensive experimental evaluation, confirming that both types of information—but especially contextual features—play an important role.



CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Discourse, dialogue and pragmatics**;

Additional Key Words and Phrases: Fact-checking, discourse, community question-answering

**ACM Reference format:**
Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic Fact-Checking Using Context and Discourse Information. *J. Data and Information Quality* 11, 3, Article 12 (May 2019), 27 pages.
https://doi.org/10.1145/3297722

Work conducted while L. Màrquez was at QCRI, HBKU.
Authors' addresses: P. Atanasova, Sofia University "St. Kliment Ohridski," 5, James Bourchier Blvd., 1164, Sofia, Bulgaria; email: pepa.k.gencheva@gmail.com; P. Nakov and A. Barrón-Cedeño, Qatar Computing Research Institute, HBKU, HBKU Research Complex B1, P.O. Box 5825, Doha, Qatar; emails: pnakov@qf.org.qa, albarron@[hbku.edu.qa|gmail.com]; L. Màrquez, Amazon, Barcelona, Spain; email: lluismv@amazon.com; G. Karadzhov and T. Mihaylova, Sofia University "St. Kliment Ohridski," 5, James Bourchier Blvd., 11641, Sofia, Bulgaria; emails: {georgi.m.karadjov, tsvetomila.mihaylova}@gmail.com; M. Mohtarami and J. Glass, Massachusetts Institute of Technology, MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA; emails: mitra@csail.mit.edu, glass@mit.edu.

1936-1955/2019/05-ART12 $15.00
https://doi.org/10.1145/3297722

## 1 INTRODUCTION

Recent years have seen the proliferation of deceptive information online. With the increasing necessity to validate information from the Internet, *automatic fact-checking* has emerged as an important research topic. Fact-checking is at the core of multiple applications, e.g., discovery of fake news (Lazer et al. 2018), rumor detection in social media (Vosoughi et al. 2018), information verification in question-answering systems (Mihaylova et al. 2018), detection of information manipulation agents (Chen et al. 2013; Darwish et al. 2017; Mihaylov et al. 2015b), and assistive technologies for investigative journalism (Hassan et al. 2015). It touches many aspects, such as credibility of users and sources, information veracity, information verification, and linguistic aspects of deceptive language. There has been work on automatic claim identification (Hassan et al. 2015, 2016) and also on checking the factuality/credibility of a claim of a news article or of an information source (Ba et al. 2016; Castillo et al. 2011; Hardalov et al. 2016; Karadzhov et al. 2017a, 2017b; Ma et al. 2016; Nakov et al. 2017b; Rashkin et al. 2017; Zubiaga et al. 2016). In general, previous work has not paid much attention to explicitly modeling contextual information and linguistic properties of the discourse to identify and verify claims, with some rare recent exceptions (Gencheva et al. 2017; Popat et al. 2017).

In this article, we focus on studying the role of *contextual information* and *discourse* that provide important information typically not included in the usual feature sets, which are mostly based on properties of the target claim and its similarity to a set of validation documents or snippets. In particular, we focus on the following tasks:

*Check-worthy claim identification.* We address the automatic identification of claims in political debates that a journalist should fact-check. In this case, the text is dialog-style: with long turns by the candidates and orchestrated by a moderator around particular topics. Journalists had to challenge the veracity of claims in the 2016 US presidential campaign, and this was particularly challenging during the debates, as a journalist had to prioritize which claims to fact-check first. Thus, we developed a model that ranks the claims by their check-worthiness.

*Answers fact-checking.* We address the automatic verification of answers in community-driven Web forums (e.g., Quora, StackOverflow). The text is thread-style but is subject to potential dialogues: a user posts a question and others post potential answers. That is, the answers are verified in the context of discussion threads in a forum and are also interpreted in the context of an initial question. Here, we deal with social media content. The text is noisier and the information being shared is not always factual, mainly due to misunderstanding, ignorance, or maliciousness of the responder.

We run extensive experiments for both tasks by training and applying classifiers based on neural networks, kernel-based support vector machines, and combinations thereof. The results confirm that the contextual and the discourse information are crucial to boost the models and to achieve state-of-the-art results for both tasks.[1] In the former task, using context yields 4.2 MAP points of absolute improvement, while using discourse information adds 1.5 MAP absolute points; in the latter task, considering the discourse and the contextual information improves the performance by a total of 4.5 MAP absolute points.

The rest of this article is organized as follows: Section 2 describes our supervised approach to predicting the check-worthiness of text fragments with focus on political debates. Section 3 presents our approach to verifying the factuality of the answers in a community question-answering forum. Section 4 provides a more qualitative analysis of the outcome of all our

[1]We make available the datasets and source code for both tasks: https://github.com/pgencheva/claim-rank and https://github.com/qcri/QLFactChecking.

Table 1. Number of Annotations in Each Medium for the 1st, 2nd, and 3rd Presidential and the Vice-presidential Debates

| Medium | 1st | 2nd | VP | 3rd | Total |
|---|---|---|---|---|---|
| ABC News | 35 | 50 | 29 | 28 | 142 |
| Chicago Tribune | 30 | 29 | 31 | 38 | 128 |
| CNN | 46 | 30 | 37 | 60 | 173 |
| FactCheck.org | 15 | 45 | 47 | 60 | 167 |
| NPR | 99 | 92 | 91 | 89 | 371 |
| PolitiFact | 74 | 62 | 60 | 57 | 253 |
| The Guardian | 27 | 39 | 54 | 72 | 192 |
| The New York Times | 26 | 25 | 46 | 52 | 149 |
| The Washington Post | 26 | 19 | 33 | 17 | 95 |
| **Total annotations** | 378 | 391 | 428 | 473 | 1,670 |
| **Annotated sentences** | 218 | 235 | 183 | 244 | 880 |

The last row shows the number of annotated sentences that become the positive examples in the CW-USPD-2016 dataset.

experiments. Section 5 discusses related work. Section 6 presents the conclusions and the lessons learned, and further outlines some possible directions for future research.

## 2 CLAIM IDENTIFICATION

In this section, we focus on the problem of automatically identifying which claims in a given document are most check-worthy and thus should be prioritized for fact-checking. We focus on how contextual and discourse information can help in this task. We further study how to learn from multiple sources simultaneously (e.g., PolitiFact, FactCheck, ABC), with the objective of mimicking the selection strategies of one particular target source; we do this in a multi-task learning setup.

### 2.1 Data

We used the CW-USPD-2016 dataset, which is centered around political debates (Gencheva et al. 2017). It contains four transcripts of the 2016 US Presidential election debates: one vice-presidential and three presidential. Each debate is annotated at the sentence level as *check-worthy* or not, but the sentences are kept in the context of the full debate, including metadata about the speaker, speaker turns, and system messages about the public reaction. The annotations were derived using publicly available manual analysis of these debates by nine reputable fact-checking sources, shown in Table 1. This analysis was converted into a binary annotation: whether a particular sentence was annotated for factuality by a given source. Whenever one or more annotations were about part of a sentence, the entire sentence was selected, and when an annotation spanned over multiple sentences, each of them was selected. The dataset with the four debates contains 5,415 sentences, out of which 880 are positive examples (i.e., selected for fact-checking by at least one of the sources). Table 2 presents an excerpt of this corpus.

Note that the investigative journalists did not select the check-worthy claims in isolation, ignoring the context. Our analysis shows that these include claims that were highly disputed during the debate, that were relevant to the topic introduced by the moderator, and so on. We will make use of these contextual dependencies below.

### 2.2 Modeling Context and Discourse

We developed a rich input representation to model and to predict the *check-worthiness* of a sentence. In particular, we included a variety of contextual and discourse-based features. They

Table 2. Excerpt from the Transcript of the First US Presidential Debate in 2016, Annotated by Nine Sources: Chicago Tribune (CT), ABC News, CNN, Washington Post (WP), NPR, PolitiFact (PF), The Guardian (TG), The New York Times (NYT), and Factcheck.org (FC)

| Speaker | Text | Annotation Sources | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CT | ABC | CNN | WP | NPR | PF | TG | NYT | FC | **All** | **Check?** |
| Clinton: | 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. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | No |
| Clinton: | 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. | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 6 | Yes |
| Clinton: | 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. | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | Yes |
| Clinton: | Take clean energy. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | No |
| Clinton: | Some country is going to be the clean- energy superpower of the 21st century. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | No |
| Clinton: | Donald thinks that climate change is a hoax perpetrated by the Chinese. | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 6 | Yes |
| Clinton: | I think it's real. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | No |
| Trump: | I did not. | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 5 | Yes |

Whether the media fact-checked the claim or not is indicated by a 1 or 0, respectively. The total number of sources that annotated an example is shown in column "All." Column "Check?" indicates the class label, i.e., whether the example is check-worthy or not. The positive examples are also highlighted in blue.

characterize the sentence in the context of the full *segment* by the same speaker, sometimes also looking at the previous and the following segments. We define a *segment* as a maximal set of consecutive sentences by the same speaker without intervention by another speaker or the moderator, i.e., a *turn*. We start by describing these context-based features, which are the focus of attention of this work.

*2.2.1 Position (*3 Features*).* A sentence on the boundaries of a speaker's segment could contain a reaction to another statement or could provoke a reaction, which in turn could signal a check-worthy claim. Thus, we added information about the position of the target sentence in its segment: whether it is first/last, as well as its reciprocal rank in the list of sentences in that segment.

*2.2.2 Segment Sizes (*3 Features*).* The size of the segment belonging to one speaker might indicate whether the target sentence is part of a long speech, makes a short comment, or is in the middle of a discussion with lots of interruptions. The size of the previous and of the next segments

is also important in modeling the dialogue flow. Thus, we include three features with the size of the previous, the current, and the next segments.

*2.2.3 Metadata (*8 Features*).* Check-worthy claims often contain accusations about the opponents, as the example below shows (from the second presidential debate):

Trump: **Hillary Clinton** attacked those same women and attacked them viciously.
...
Clinton: They're doing it to try to influence the election for **Donald Trump**.

Thus, we use a feature that indicates whether the target sentence mentions the name of the opponent, whether the speaker is the moderator, and also who is speaking (three features). We further use three binary features, indicating whether the target sentence is followed by a system message: *applause*, *laugh*, or *cross-talk*.

*2.2.4 Topics (*303 Features*).* Some topics are more likely to be associated with check-worthy claims, and thus we have features modeling the topics in the target sentence as well as in the surrounding context. We trained a Latent Dirichlet Allocation (LDA) topic model (Blei et al. 2003) on all political speeches and debates in *The American Presidency Project*[2] using all US presidential debates in the 2007–2016 period.[3] We had 300 topics, and we used the distribution over the topics as a representation for the target sentence. We further modeled the context using cosines with such representations for the previous, the current, and the next segment.

*2.2.5 Embeddings (*303 Features*).* We also modeled semantics using word embeddings. We used the pre-trained 300-dimensional Google News word embeddings by Mikolov et al. (2013a) to compute an average embedding vector for the target sentence, and we used the 300 dimensions of that vector. We also modeled the context as the cosine between that vector and the vectors for three segments: the previous, the current, and the following one.

*2.2.6 Contradictions (*5 features*).* Many claims selected for fact-checking contain contradictions to what has been said earlier, as in the example below (from the third presidential debate):

Clinton: [...] about a potential nuclear competition in Asia, you said, you know, go ahead, enjoy yourselves, folks.
Trump: **I didn't say** nuclear.

We model this by counting the negations in the target sentence as found in a dictionary of negation cues such as *not*, *didn't*, and *never*. We further model the context as the number of such cues in the two neighboring sentences from the same segment and the two neighboring segments.

*2.2.7 Similarity of the Sentence to Known Positive/Negative Examples (*3 Features*).* We used three more features that measure the similarity of the target sentence to other known examples. The first one computes the maximum over the training sentences of the number of matching words between the target and the training sentence, which is further multiplied by −1 if the latter was not check-worthy. We also used another version of the feature, where we multiplied it by 0 if the speakers were different. A third version took as a training set all claims checked by *PolitiFact*[4] (excluding the target sentence).

[2] http://www.presidency.ucsb.edu/debates.php.
[3] https://github.com/paigecm/2016-campaign.
[4] http://www.politifact.com/.

*2.2.8 Discourse (*20 Features*).* We saw above that contradiction can signal the presence of check-worthy claims, and contradiction can be expressed by a discourse relation such as Contrast. As other discourse relations such as Background, Cause, and Elaboration can also be useful, we used a discourse parser (Joty et al. 2015) to parse the entire segment. This parser follows the Rhetorical Structure Theory (RST). It produces a hierarchical representation of the discourse by linking first the elementary discourse units with binary discourse relations (indicating also which unit is the *nucleus* and which is the *satellite*), and building up the tree by connecting with the same type of discourse relations the more general cross-sentence nodes until a root node covers all the text. From this tree, we focused on the direct relationship between the target sentence and the other sentences in its segment; this gave rise to 18 contextual indicator features. We further analyzed the internal structure of the target sentence—how many nuclei and how many satellites it contains—which gave rise to two sentence-level features.

### 2.3 Other Features

*2.3.1 ClaimBuster-based (*1,045 Core Features*).* In order to be able to compare our model and features directly to the previous state-of-the-art, we re-implemented, to the best of our ability, the sentence-level features of *ClaimBuster* (Hassan et al. 2015), namely TF-IDF-weighted bag of words (998 features), part-of-speech tags (25 features), named entities as recognized by *Alchemy API*[5] (20 features), sentiment score from *Alchemy API* (1 feature), and number of tokens in the target sentence (1 feature). Apart from providing means of comparison to the state-of-the-art, these features also make a solid contribution to our final system for check-worthiness estimation. However, note that we did not have access to the training data of *ClaimBuster*, which is not publicly available, and we thus train on our own dataset.

*2.3.2 Sentiment (*2 Features*).* Some sentences are highly negative, which can signal the presence of an interesting claim to check, as the two following example sentences show (from the first and the second presidential debates):

Trump: Murders are up.
Clinton: Bullying is up.

We used the NRC sentiment lexicon (Mohammad and Turney 2013) as a source of words and *n*-grams with positive/negative sentiment, and we counted the number of positive and of negative words in the target sentence. These features are different from those in *ClaimBuster*, where these lexicons were not used.

*2.3.3 Named Entities (NE) (*1 Feature*).* Sentences that contain named entity mentions are more likely to contain a claim that is worth fact-checking, as they discuss particular people, organizations, and locations. Thus, we have a feature that counts the number of named entities in the target sentence; we use the *NLTK toolkit* for named entity recognition (Loper and Bird 2002). Unlike the *ClaimBuster* features above, here we only have one feature; we also use a different toolkit for named entity recognition.

*2.3.4 Linguistic Features (13 Features).* We use as features the presence and the frequency of occurrence of linguistic markers such as *factives* and *assertives* from Hooper (1974), *implicatives* from Karttunen (1971), *hedges* from Hyland (2005), *Wiki-bias* terms from Recasens et al. (2013), *subjectivity* cues from Riloff and Wiebe (2003), and *sentiment* cues from Liu et al. (2005).[6] We

[5] http://www.ibm.com/watson/alchemy-api.html.
[6] Most of these bias cues can be found at http://people.mpi-sws.org/~cristian/Biased_language.html.

Table 3. Some Cues for Various Bias Types

| Bias Type | Sample Cues | Bias Type | Sample Cues |
|---|---|---|---|
| Factives | realize, know, discover, learn | Modals | can, must, will, shall |
| Implicatives | cause, manage, hesitate, neglect | Negations | neither, without, against, never, none |
| Assertives | think, believe, imagine, guarantee | Strong-subj | admire, afraid, agreeably, apologist |
| Hedges | approximately, estimate, essentially | Weak-subj | abandon, adaptive, champ, consume |
| Report-verbs | argue, admit, confirm, express | Positives | accurate, achievements, affirm |
| Wiki-bias | capture, create, demand, follow | Negatives | abnormal, bankrupt, cheat, conflicts |

compute a feature vector according to Equation (1), where for each bias type $B_i$ and answer $A_j$, the frequency of the cues for $B_i$ in $A_j$ is computed and then normalized by the total number of words in $A_j$:

$$B_i(A_j) = \frac{\sum_{cue \in B_i} count(cue, A_j)}{\sum_{w_k \in A_j} count(w_k, A_j)}. \quad (1)$$

Below, we describe these cues in more detail (see Table 3 for examples).

- *Factives* (*1 feature*) (Hooper 1974) are verbs that imply the veracity of their complement clause. In *E1*, *know* suggests that "they will open a second school …" and "they provide a qualified french education …" are factually true statements.
  *E1:* ***know*** that they ***will*** open a second school; and they are a ***nice*** french school…I ***know*** that they ***provide*** a ***qualified*** french education and add with that the history and arabic language to be adapted to the qatar. I ***think*** that's an ***interesting*** addition.
- *Assertives* (*1 feature*) (Hooper 1974) are verbs that imply the veracity of their complement clause with a level of certainty. E.g., in *E1*, *think* indicates some uncertainty, while verbs like *claim* cast doubt on the certainty of their complement clause.
- *Implicatives* (*1 feature*) (Karttunen 1971) are verbs that imply the (un)truthfulness of their complement clause, e.g., *decline* and *succeed*.
- *Hedges* (*1 feature*) (Hyland 2005) reduce the person's commitment to the truth, e.g., *may* and *possibly*.
- *Reporting verbs* (*1 feature*) are used to report a statement from a source, e.g., *argue* and *express*.
- *Wiki-bias cues* (*1 feature*) (Recasens et al. 2013) are extracted from the NPOV corpus from Wikipedia and cover bias cues (e.g., *provide* in *E1*) and controversial words, such as *abortion* and *execute*. These words are not available in none of the other bias lexicons.
- *Modals* (*1 feature*) are used to change the certainty of the statement (e.g., *will* or *can*), make an offer (e.g., *shall*), ask permission (e.g., *may*), or express an obligation or necessity (e.g., *must*).
- *Negations* (*1 feature*) are used to deny or make negative statements such as *no*, *never*.
- *Subjectivity cues* (*2 features*) (Riloff and Wiebe 2003) are used when expressing personal opinions and feelings. There are *strong* and *weak* cues, e.g., in *E1*, *nice* and *interesting* are *strong*, while *qualified* is *weak*.
- *Sentiment cues* (*2 features*). We use *positive* and *negatives* sentiment cues (Liu et al. 2005) to model the attitude, thought, and emotions of the speaker. In *E1*, *nice*, *interesting*, and *qualified* are positive cues.

The above bias and subjectivity cues are mostly single words. Sometimes a multi-word cue (e.g., "we can guarantee") can be a stronger signal for a user's certainty/uncertainty in their answers. We thus further generate multi-word cues *(1 feature)* by combining *implicative*, *assertive*, *factive,* and *report* verbs with first-person pronouns (*I/we*), *modals* and strong subjective *adverbs*, e.g., *I/we+verb* (e.g. "I believe"), *I/we+adverb+verb* (e.g., "I certainly know"), *I/we+modal+verb* (e.g., "we could figure out"), and *I/we+modal+adverb+verb* (e.g., "we can obviously see").

*2.3.5 Tense (*1 Feature*).* Most of the check-worthy claims mention past events. In order to detect when the speaker is making a reference to the past or is talking about his/her future vision and plans, we include a feature with three values—indicating whether the text is in past, present, or future tense. The feature is extracted in a simplified fashion from the verbal expressions, using POS tags and a list of auxiliary phrases. In particular, we consider a sentence to be in the past tense if it contains a past verb (*VBD*), and in the future tense if it contains *will* or *have to*; otherwise, we assume it to be in the present tense.

*2.3.6 Length (*1 Feature*).* Shorter sentences are generally less likely to contain a check-worthy claim.[7] Thus, we have a feature for the length of the sentence in terms of characters. Note that this feature was not part of the *ClaimBuster* features, as their length was modeled in terms of tokens, but here we do so using characters.

### 2.4 Experiments

*Learning Algorithm.* We used a feed-forward neural network (FNN) with two hidden layers (with 200 and 50 neurons, respectively) and a softmax output unit for the binary classification.[8] We used ReLU (Glorot et al. 2011) as the activation function and we trained the network with Stochastic Gradient Descent (LeCun et al. 1998) for 300 epochs with a batch size of 550. We set the L2 regularization to 0.0001, and we kept a constant learning rate of 0.04. We further enhanced the learning process by using a Nesterov's momentum (Sutskever et al. 2013) of 0.9.

*Setting.* We trained the models to classify sentences as positive if *one or more media* had fact-checked a claim inside the target sentence, and negative otherwise. We then used the classifier scores to rank the sentences with respect to *check-worthiness*.[9] We tuned the parameters and we evaluated the performance using 4-fold cross-validation, using each of the four debates in turn for testing while training on the remaining three.

*Implementation Details.* We used `gensim` (Řehuřek and Sojka 2010) for LDA and word embeddings, `NLTK` (Loper and Bird 2002) for NER and POS tagging, and `scikit-learn` (Buitinck et al. 2013) for deep learning.

*Evaluation.* We use ranking measures such as *Precision at k* (*P@k*) and *Mean Average Precision* (MAP). As Table 1 shows, most media rarely check more than 50 claims per debate, which means that there is no need to fact-check more than 50 sentences. Thus, we report *P@k* for $k \in \{5, 10, 20, 50\}$.[10] MAP is the mean of the Average Precision across the four debates. Finally, we also measure the recall at the *R*th position of returned sentences for each debate, where *R* is

[7] One notable exception are short sentences with negations, e.g., *Wrong.*, *Nonsense.*, and so on.

[8] Previous work (Gencheva et al. 2017) showed that the neural network performs better on the task than support vector machine classifiers.

[9] We also tried using ordinal regression, and SVM-perf (an instantiation of SVM-struct), to directly optimize precision, but they performed worse.

[10] Note that as far as the difference between the P@k metrics (especially between 5 and 10) is in terms of a few sentences, the deviation between them can seem large, while caused by a few correctly/wrongly classified sentences.

Table 4. Overall Results for Check-worthy Claims Identification, Focusing on the Impact of the Contextual and Discourse Features

| **Our System** | **MAP** | **R-Pr** | **P@5** | **P@10** | **P@20** | **P@50** |
|---|---|---|---|---|---|---|
| All features | **0.427** | **0.432** | **0.800** | **0.725** | **0.713** | **0.600** |
| All\discourse | 0.412 | 0.431 | 0.800 | 0.700 | 0.685 | 0.550 |
| All\context | 0.385 | 0.390 | 0.550 | 0.500 | 0.550 | 0.540 |
| Only context+discourse | 0.317 | 0.404 | 0.725 | 0.563 | 0.465 | 0.465 |
| **Reference systems** | | | | | | |
| *Random* | 0.164 | 0.007 | 0.200 | 0.125 | 0.138 | 0.160 |
| *TF-IDF* | 0.314 | 0.333 | 0.550 | 0.475 | 0.413 | 0.360 |
| *Claimbuster–Platform* | 0.317 | 0.349 | 0.500 | 0.550 | 0.488 | 0.405 |
| *Claimbuster–Features* | 0.357 | 0.379 | 0.500 | 0.550 | 0.550 | 0.510 |

the number of relevant documents for that debate and the metric is known as *R*-Precision (*R*-Pr). As with MAP, we provide the average across the four debates.

*Results.* Table 4 shows all the results of our claim ranking system with several feature variants. In order to put the numbers in perspective, we also show the results for four increasingly competitive baselines ("Reference Systems"). The first one is a random baseline. It is then followed by an SVM classifier based on a bag-of-words representation with TF-IDF weights estimated on the training data. Then come two versions of the *ClaimBuster* system: Claimbuster–Platform refers to the performance of *ClaimBuster* using the scores obtained from their online demo,[11] which we accessed on December 20, 2016, and Claimbuster–Features is our re-implementations of *ClaimBuster* using our FNN classifiers trained on our dataset with their features.

We can see that our system with all features outperforms all reference systems by a large margin for all metrics. The two versions of ClaimBuster also outperform the TF-IDF baseline on most measures. Moreover, our re-implementation of *ClaimBuster* is better than the online platform, especially in terms of MAP. This is expected, as their system is trained on a different dataset and it may suffer from testing on slightly out-of-domain data. Our advantage with respect to ClaimBuster implies that the extra information coded in our model, mainly more contextual, structural, and linguistic features, has an important contribution to the final performance.

Rows 2–4 in Table 4 show the effect of the discourse and of the contextual features implemented in our system. The contextual features have a major impact on performance: excluding them yields major drops for all measures, e.g., MAP drops from 0.427 to 0.385, and P@5 drops from 0.800 to 0.550. The discourse features also have an impact, although it is smaller. The most noticeable difference is in the quality at the lower positions in the rank, e.g., P@5 does not vary when removing discourse features, but P@10, P@20, and P@50 all drop by 2.5 to 5 percent points. Finally, row 4 in the table shows that contextual+discourse features alone already yield a competitive system, performing about the same as *Claimbuster–Platform* (which uses no contextual features at all). In Section 4, we will present a further qualitative description of the results including some examples.

### 2.5 Multi-task Learning Experiments

Unlike the above single-source approaches, in this subsection, we explore a multi-source neural network framework in which we try to predict the selections of each and every fact-checking organization simultaneously. We show that, even when the goal is to mimic the selection strategy

[11] http://idir-server2.uta.edu/claimbuster/demo.

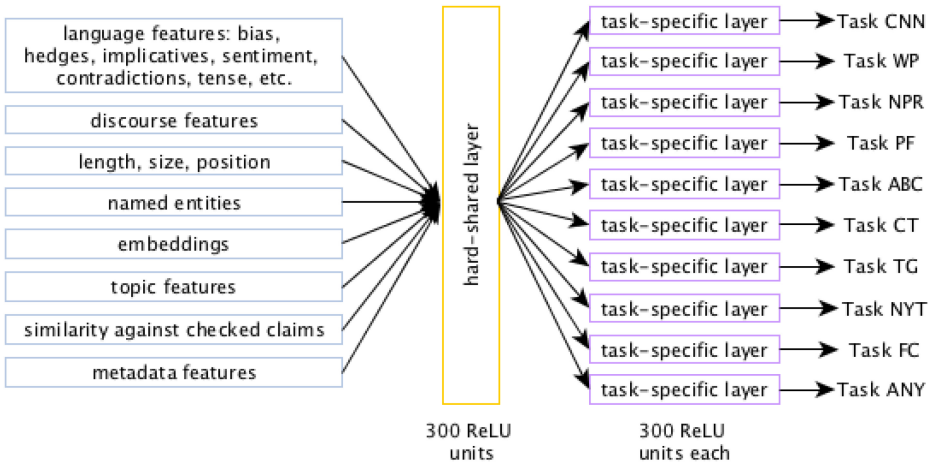



Fig. 1. The architecture of the full neural multi-source learning model, predicting the selection choices of each of the nine individual sources (tasks) and of one cumulative source: *task ANY*.

of one particular fact-checking organization, it is beneficial to leverage on the selection choices by multiple such organizations.

*Setting.* We approach the task of check-worthiness prediction using the same features, while at the same time modeling the problem as multi-task learning, using different sources of annotation over the same training dataset. As a result, we can learn to mimic the selection strategy of each and every one of these individual sources. As we have explained above, in our dataset the individual judgments come from nine independent fact-checking organizations, and we thus predict the selection choices of each of them in isolation plus a collective label *ANY*, which indicates whether at least one source would judge that claim as check-worthy.

*Architecture.* Figure 1 illustrates the architecture of the full neural multi-source learning model, which predicts the selection choices of each of the nine individual sources (tasks) and of the special cumulative source: *task ANY*. There is a hidden layer (of size 300) that is shared between all ten tasks. Then, each task has its own task-specific hidden layer (each of size 300). 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. During training, each task modifies the weights of both its own task-specific layer and of the shared layer. For our neural network architecture, we used ReLU units, Stochastic Gradient Descent with Nesterov momentum of 0.7, iterating for 100 epochs with batches of size 500 and a learning rate of 0.08.

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. In particular, it has been shown that the risk of overfitting the shared parameters in the hidden layer is an order $n$ smaller than overfitting the task-specific parameters in the output layers, where $n$ is the number of tasks at hand (Baxter 1997). The input to our neural network consists of the various domain-specific features that have been previously described.

*Implementation Details.* We implemented the neural network using Keras. We tried adding more shared and task-specific layers as well as having some task-specific layers linked directly to the input, but we eventually settled on the architecture in Figure 1. We also tried to optimize directly

Table 5. Evaluation Results for Each of the Nine Fact-checking Organizations as a Target to Mimic

| Model | MAP | R-Pr | P@5 | P@10 | P@20 | P@50 |
|---|---|---|---|---|---|---|
| **ABC** | | | | | | |
| *singleton* | 0.097 | 0.112 | 0.250 | 0.175 | 0.162 | 0.100 |
| ***multi*** | **0.119** | **0.157** | **0.333** | **0.225** | **0.217** | **0.122** |
| ***multi+any*** | **0.118** | **0.160** | **0.300** | **0.233** | **0.229** | **0.132** |
| **The Washington Post (WP)** | | | | | | |
| *singleton* | 0.106 | 0.110 | 0.150 | 0.100 | 0.112 | 0.110 |
| ***multi*** | **0.127** | **0.127** | **0.350** | **0.233** | **0.162** | **0.123** |
| ***multi+any*** | **0.130** | **0.129** | **0.350** | **0.250** | **0.171** | ***0.110*** |
| **CNN** | | | | | | |
| *singleton* | 0.087 | 0.091 | 0.250 | 0.150 | 0.121 | 0.090 |
| ***multi*** | **0.113** | **0.132** | ***0.250*** | **0.208** | **0.183** | **0.140** |
| ***multi+any*** | **0.109** | **0.126** | 0.167 | **0.200** | **0.167** | **0.128** |
| **FactCheck (FC)** | | | | | | |
| *singleton* | 0.084 | 0.114 | 0.117 | 0.125 | 0.088 | 0.100 |
| ***multi*** | **0.105** | **0.136** | **0.250** | **0.175** | **0.146** | **0.118** |
| ***multi+any*** | **0.117** | 0.110 | **0.333** | **0.242** | **0.196** | **0.107** |
| **PolitiFact** | | | | | | |
| *singleton* | 0.201 | 0.278 | 0.250 | 0.250 | 0.262 | 0.262 |
| ***multi*** | **0.209** | 0.258 | **0.400** | **0.367** | **0.317** | **0.270** |
| ***multi+any*** | **0.210** | 0.252 | **0.500** | **0.350** | **0.333** | **0.272** |

| Model | MAP | R-Pr | P@5 | P@10 | P@20 | P@50 |
|---|---|---|---|---|---|---|
| **NPR** | | | | | | |
| *singleton* | 0.175 | 0.195 | 0.250 | 0.250 | 0.283 | 0.228 |
| ***multi*** | **0.186** | **0.210** | **0.333** | **0.342** | **0.300** | **0.245** |
| ***multi+any*** | **0.180** | **0.207** | **0.333** | **0.283** | 0.250 | 0.227 |
| **The Guardian (TG)** | | | | | | |
| *singleton* | 0.127 | 0.174 | 0.200 | 0.150 | 0.196 | 0.178 |
| ***multi*** | **0.133** | **0.199** | 0.183 | **0.175** | **0.192** | **0.193** |
| ***multi+any*** | **0.130** | 0.159 | **0.217** | **0.175** | **0.200** | 0.167 |
| **Chicago Tribune (CT)** | | | | | | |
| *singleton* | 0.079 | 0.110 | 0.100 | 0.100 | 0.125 | 0.075 |
| ***multi*** | **0.081** | 0.090 | ***0.100*** | **0.133** | 0.104 | **0.082** |
| ***multi+any*** | **0.087** | 0.087 | **0.133** | ***0.100*** | 0.108 | **0.093** |
| **The New York Times (NYT)** | | | | | | |
| *singleton* | 0.187 | 0.221 | 0.350 | 0.325 | 0.238 | 0.192 |
| ***multi*** | 0.150 | 0.213 | 0.233 | 0.200 | 0.196 | 0.180 |
| ***multi+any*** | 0.147 | 0.197 | 0.200 | 0.167 | 0.158 | 0.162 |

Shown are results for single-source baselines and for multi-task learning. The improvements over the corresponding baselines are marked in bold.

for average precision and adding loss weights to *task ANY*, but using the standard binary cross-entropy loss yielded the best results.

*Results.* As before, we perform 4-fold cross-validation, where each time we leave one debate out for testing. Moreover, to stabilize the results, we repeat each experiment three times with different random seeds, and we report the average over these three reruns.[12] We should note that in most cases this was not really needed, as the standard deviation for the reruns was generally tiny: 0.001 or less, absolute.

Table 5 presents the results, with all evaluation metrics, when predicting each of the nine sources. We experiment with three different configurations of the model described in the previous section. All of them aim at learning to mimic the selection choices by one single fact-checking organization (source). The first one is a single-task baseline *singleton* where a separate neural network is trained for each source. The other two are multi-task learning configurations: *multi* trains to predict labels for each of the nine tasks (one for each fact-checker); and *multi+any* trains to predict labels for each of the nine tasks (one for each fact-checker), and also for *task ANY* (as shown in Figure 1). We can see in Table 5 that, for most of the sources, multi-task learning improves over the single-source system. The results of the multi-task variations that improve over the single baseline are boldfaced in the table. The improvements are consistent across evaluation metrics and

[12]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).

Table 6. Evaluation Results Averaged Over Nine Fact-checking Organizations

| Model | MAP | R-Pr | P@5 | P@10 | P@20 | P@50 |
|---|---|---|---|---|---|---|
| CB online | 0.090 | 0.138 | 0.144 | 0.143 | 0.121 | 0.117 |
| singletonG | 0.120 | 0.142 | 0.228 | 0.206 | 0.179 | 0.137 |
| *any* | 0.128 | 0.225 | 0.194 | 0.186 | 0.178 | 0.153 |
| *singleton (embed.)* | 0.058 | 0.065 | 0.055 | 0.055 | 0.068 | 0.072 |
| singleton CB | 0.072 | 0.077 | 0.106 | 0.076 | 0.081 | 0.079 |
| *singleton* | 0.127 | 0.156 | 0.213 | 0.181 | 0.176 | 0.148 |
| ***multi*** | **0.136** | **0.169** | **0.270** | **0.229** | **0.202** | **0.164** |
| ***multi+any*** | **0.136** | **0.159** | **0.281** | **0.222** | **0.201** | **0.155** |
| *any* | 0.125 | 0.153 | 0.204 | 0.197 | 0.175 | 0.153 |
| *singleton+any* | **0.130** | 0.153 | **0.237** | **0.220** | **0.184** | ***0.148*** |

The improvements over *singleton* are in bold.

vary largely depending on the source and the metric. 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.

The first three rows of Table 6 present the same results but averaged over the nine sources. Again, we can see that multi-task learning yields sizable improvement over the single-task learning baseline for all evaluation measures. Another conclusion that can be drawn is that including the task *any* 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 distinct sources only. The last two rows in Table 6 present two additional variants of the model: the single-task learning *any* system, which is trained on the union of the selected sentences by all nine fact-checkers to predict the target fact-checker only; and the system *singleton+any*, which predicts labels for two tasks: (*i*) for the target fact-checker and (*ii*) for *task ANY*.

We can see that the model *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 *task ANY* is not enough.

## 3 FACT-CHECKING

With the ever-growing amount of unreliable content online, veracity will almost certainly become an important component of question-answering systems in the future. In this section, we focus on fact-checking in the context of community question answering (cQA), i.e., predicting whether an answer to a given question is likely to be true. This aspect has been ignored, e.g., in recent cQA tasks at NTCIR and SemEval (Ishikawa et al. 2010; Nakov et al. 2017a, 2015, 2016), where an answer is considered as GOOD if it tries to address the question, irrespective of its veracity. Yet, veracity is an important aspect, as high-quality automatic fact-checking can offer a better experience to the users of cQA systems; e.g., a possible application scenario would be that in which the user could be presented with a ranking of all good answers accompanied by veracity scores, where low scores would warn her not to completely trust the answer or to double-check it.

$q$: If wife is under her husband's sponsorship and is willing to come Qatar on visit, how long she can stay after extending the visa every month? I have heard it's not possible to extend visit visa more than 6 months? ...

$a_1$: Maximum period is 9 Months....
$a_2$: 6 months maximum
$a_3$: This has been answered in QL so many times. Please do search for information regarding this. BTW answer is 6 months.

Fig. 2. Example from the Qatar Living forum.

Table 7. Distribution of Factuality Labels in the Annotated Questions (1,357 in total)

| Label | #Questions | Example |
|---|---|---|
| Factual | 373 | What is Ooredoo customer service number? |
| Opinion | 689 | Can anyone recommend a good Vet in Doha? |
| Socializing | 295 | What was your first car? |

Figure 2 presents an excerpt of an example from the Qatar Living forum, with one question ($q$) and three plausible answers ($a_1 - a_3$) selected from a longer thread. According to the SemEval-2016 Task 3 annotation instructions (Nakov et al. 2016), all three answers are considered Good since they address the question. Nevertheless, $a_1$ contains false information, while $a_2$ and $a_3$ are true,[13] as can be checked on an official governmental website.[14]

### 3.1 Data

We use the CQA-QL-FACT dataset, which stresses the difference between (*i*) distinguishing a good vs. a bad answer, and (*ii*) distinguishing between a factually true vs. a factually false one. We added the factuality annotations on top of the CQA-QL-2016 dataset from the SemEval-2016 Task 3 on community Question Answering (Nakov et al. 2016). In CQA-QL-2016, the data is organized in question–answer threads extracted from the Qatar Living forum. Each question has a subject, a body, and metadata: ID, category (e.g., *Computers and Internet*, *Education*, and *Moving to Qatar*), date and time of posting, and user name.

First, we annotated the questions using the following labels:

- Factual: The question is asking for factual information, which can be answered by checking various information sources, and it is not ambiguous.
- Opinion: The question asks for an opinion or advice, not for a fact.
- Socializing: Not a real question, but rather socializing/chatting. This can also mean expressing an opinion or sharing some information without really asking anything of general interest.

We annotated 1,982 questions with the above factuality labels. We ended up with 625 instances that contain multiple questions, which we excluded from further analysis. Table 7 shows the annotation results for the remaining 1,357 questions, including examples.

[13] One could also guess that answers $a_2$ and $a_3$ are more likely to be true from the fact that the *6 months* answer fragment appears many times in the current thread (it also happens to appear more often in related threads as well). While these observations serve as the basis for useful features for classification, the real verification for a gold-standard annotation requires finding support from a reliable external information source: in this case, an official government information portal.
[14] https://www.moi.gov.qa/site/english/departments/PassportDept/news/2011/01/03/23385.html.

Next, we annotated for veracity the answers to the factual questions. We only annotated the originally judged as Good answers (ignoring both Bad and Potentially Useful ), and we used the following labels:

- Factual - True: The answer is True and this can be verified using an external resource. (*q: "I wanted to know if there were any specific shots and vaccinations I should get before coming over [to Doha]."; a: "Yes there are; though it varies depending on which country you come from. In the UK, the doctor has a list of all countries and the vaccinations needed for each."*)[15]
- Factual - False: The answer gives a factual response, but it is false. (*q: "Can I bring my pitbulls to Qatar?"; a: "Yes you can bring it but be careful this kind of dog is very dangerous."*)[16]
- Factual - Partially True: We could only verify part of the answer. (*q: "I will be relocating from the UK to Qatar [...] is there a league or TT clubs/nights in Doha?"; a: "Visit Qatar Bowling Center during thursday and friday and you'll find people playing TT there."*)[17]
- Factual - Conditionally True: The answer is True in some cases, and False in others, depending on some conditions that the answer does not mention. (*q: "My wife does not have NOC from Qatar Airways; but we are married now so can i bring her legally on my family visa as her husband?"; a: "Yes you can."*)[18]
- Factual - Responder Unsure: The person giving the answer is not sure about the veracity of his/her statement. (e.g., "*Possible only if government employed. That's what I heard.*")
- NonFactual: The answer is not factual. It could be an opinion, advice, and so on, that cannot be verified. (e.g., "*It's better to buy a new one.*")

We further discarded items whose factuality was very time-sensitive (e.g., "*It is Friday tomorrow.*"; "*It was raining last week.*"),[19] or for which the annotators were unsure.

We considered all questions from the Dev and the Test partitions of the CQA-QL-2016 dataset. We targeted very high quality, and thus we did not crowdsource the annotation, as pilot annotations showed that the task was very difficult and that it was not possible to guarantee that Turkers would do all the necessary verification, e.g., gather evidence from trusted sources. Instead, all examples were first annotated independently by four annotators, and then they discussed *each example* in detail to come up with a final label. We ended up with 249 Good answers[20] to 71 different questions, which we annotated for factuality: 128 Positive and 121 Negative examples. See Table 8 for details.

### 3.2 Modeling Context and Discourse

We model the *context* of an answer with respect to the entire answer thread in which it occurs and with respect to other high-quality posts from the entire Qatar Living forum. We further use *discourse* features as in Section 2.2.8.

[15] This can be verified at https://wwwnc.cdc.gov/travel/destinations/traveler/none/qatar.

[16] The answer is not true because pitbulls are included in the list of banned breeds in Qatar: http://canvethospital.com/pet-relocation/banned-dog-breed-list-qatar-2015/.

[17] The place has table tennis, but we do not know on which days: https://www.qatarbowlingfederation.com/bowling-center/.

[18] This answer can be true, but this depends upon some conditions: http://www.onlineqatar.com/info/dependent-family-visa.aspx.

[19] Arguably, many answers are somewhat time-sensitive, e.g., "*There is an IKEA in Doha*" is true only after IKEA opened, but not before that. In such cases, we just used the present situation as a point of reference.

[20] This is comparable in size to other fact-checking datasets, e.g., Ma et al. (2015) used 226 rumors, and Popat et al. (2016) had 100 Wiki hoaxes.

Table 8. Distribution of the Positive and the Negative Answers (i.e., the Two Classes We Predict) and of the Fine-grained Labels

| Coarse-Grained Label | Answers | Fine-Grained Label | Answers |
|---|---|---|---|
| + **Positive** | **128** | + Factual - True | 128 |
| − **Negative** | **121** | − Factual - False | 22 |
| | | − Factual - Partially True | 38 |
| | | − Factual - Conditionally True | 16 |
| | | − Factual - Responder Unsure | 26 |
| | | − NonFactual | 19 |

*3.2.1 Support from the current thread (*5 Features*).* We use the cosine similarity between an answer- and a thread-vector of all Good answers using Qatar Living embeddings. For this purpose, we use 100-dimensional in-domain word embeddings (Mihaylov and Nakov 2016a), which were trained using word2vec (Mikolov et al. 2013b) on a large dump of Qatar Living data (2M answers).[21] The idea is that if an answer is similar to other answers in the thread, it is more likely to be true. To this, we add thread-level features related to the rank of the answer in the thread: (*i*) the reciprocal rank of the answer in the thread and (*ii*) percentile of answer's rank in the thread. As there are exactly ten answers per thread in the dataset, the first answer gets the score of 1.0, the second one gets 0.9, the next one gets 0.8, and so on. We calculate these two ranking features twice: once for the full list of answers, and once for the list of good answers only.

*3.2.2 Support from all of Qatar Living (*60 Features*).* We further collect supporting evidence from all threads in the Qatar Living forum. To do this, we query a search engine, limiting the search to the forum only. See Section 3.3.3 for more detail about how the search for evidence on the Web is performed and what features are calculated.

*3.2.3 Support from high-quality posts in Qatar Living.* Among the 60K active users of the Qatar Living forum, there is a community of 38 trusted users who have written 5.2K high-quality articles on topics that attract a lot of interest, e.g., issues related to visas, work legislation, and so on. We try to verify the answers against these high-quality posts. (*i*) Since an answer can combine both relevant and irrelevant information with respect to its question, we first generate a query against a search engine for each Q&A. (*ii*) We then compute cosines between the query and the sentences in the high-quality posts, and we select the $k$-best matches. (*iii*) Finally, we compute textual entailment scores (Kouylekov and Negri 2010) for the answer given the $k$-best matches, which we then use as features. An example is shown in Table 9.

*3.2.4 Discourse Features.* We use the same discourse features as for the claim identification task (cf. Section 2.2.8).

### 3.3 Other Features

*3.3.1 Linguistic bias, subjectivity and sentiment.* Forum users, consciously or not, often put linguistic markers in their answers, which can signal the degree of the user's certainty in the veracity of what they say. We thus use the linguistic features from the previous task (see above).

*3.3.2 Credibility (*31 features*).* We use features that have been previously proposed for credibility detection (Castillo et al. 2011): number of URLs/images/emails/phone numbers; number

[21] Available at http://alt.qcri.org/semeval2016/task3/data/uploads/QL-unannotated-data-subtaskA.xml.zip.

Table 9. Sample of Sentences from High-quality Posts Automatically Extracted to Support the Answer *A*

**Question**: does anyone know if there is a french speaking nursery in doha?
**Answer**: there is a french school here. don't know the ages but my neighbor's 3 yr old goes there …
**Best Matched Sentence for Q&A:** there is a french school here.

| Post Id | sNo | R1 | R2 | Sentence |
|---|---|---|---|---|
| 35639076 | 15 | 1 | 10 | the pre-school follows the english program but also gives french and arabic lessons. |
| 32448901 | 4 | 2 | 11 | france bought the property in 1952 and since 1981 it has been home to the french institute. |
| 31704366 | 7 | 3 | 1 | they include one indian school, two french, seven following the british curriculum … |
| 27971261 | 6 | 4 | 4 | the new schools include six qatari, four indian, two british, two american and a finnish … |

*sNo* is the sentence's sequential number in the post, *R1* and *R2* are the ranks of the target sentences based on entailment and similarity, respectively.

of tokens/sentences; average number of tokens; number of positive/negative smileys; number of single/double/triple exclamation/interrogation symbols. To this set, we further add number of interrogative sentences; number of nouns/verbs/adjectives/adverbs/pronouns; and number of words that are not in word2vec's Google News vocabulary (such OOV words could signal slang, foreign language, etc.). We also use the number of first-, second-, third-person pronouns in the comments: (*i*) in absolute number, and also (*ii*) normalized by the total number of pronouns in the comment. The latter is also a feature.

*3.3.3 Support from the Web (*60 Features*).* We tried to verify whether an answer's claim is factually true by searching for supporting information on the Web. We started with the concatenation of an answer to the question that heads the respective thread. Then, following Potthast et al. (2013), we extracted nouns, verbs, and adjectives, sorted by TF-IDF (we computed IDF on the Qatar Living dump). We further extracted and added the named entities from the text and we generated a query of 5–10 words. If we did not obtain ten results, we dropped some terms and we tried again.

We automatically queried Bing and Google, and we extracted features from the resulting pages, considering Qatar-related websites only. An example is shown in Table 10. Based on the results, we calculated similarities: (*i*) cosine with TF-IDF vectors, (*ii*) cosine using Qatar Living embeddings, and (*iii*) containment (Lyon et al. 2001). We calculated these similarities between, on the one side, (*i*) the question or (*ii*) the answer or (*iii*) the question–answer pair, vs. on the other side, (*i*) the snippets or (*ii*) the web pages. To calculate the similarity to a web page, we first converted the page to a list of rolling sentence triplets, then we calculated the score of the Q/A/Q-A vs. this triplet, and finally we took the average and also the maximum similarity over these triplets. Now, as we had up to ten Web results, we further took the maximum and the average over all the above features over the returned Qatar-related pages. We created three copies of each feature, depending on whether it came from a (*i*) reputed source (e.g., news, government websites, official sites of companies, etc.), from a (*ii*) forum-type site (forums, reviews, social media), or (*iii*) from some other type of website.

*3.3.4 Embeddings (*260 features*).* Finally, we used as features the embeddings of the claim (i.e., the answer) of the best-scoring snippet and of the best-scoring sentence triplet from a web page. We calculated these embeddings using long short-term memory (LSTM) representations, which we

Table 10. Sample Snippets Returned by a Search Engine for a Given Query Generated from a Q&A Pair

| | |
|---|---|
| **Question:** | Hi; Just wanted to confirm Qatar's National Day. Is it 18th of December? Thanks. |
| **Answer:** | yes; it is 18th Dec. |
| **Query generated from Q&A:** | ``National Day'' ``Qatar'' National December Day confirm wanted |

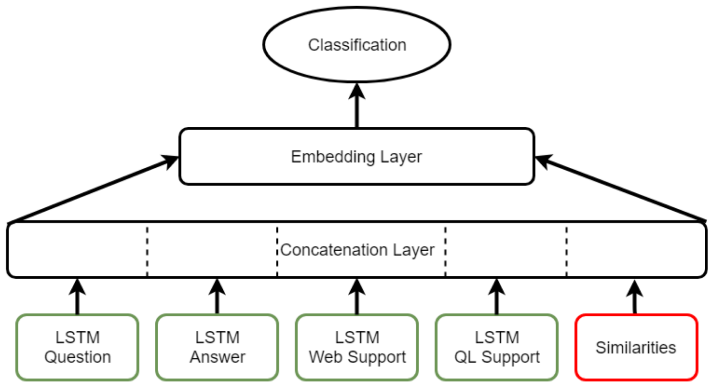
| URL | Qatar-related? | Source type | Snippet |
|---|---|---|---|
| qppstudio.net | No | Other | Public holidays and national …the world's source of Public holidays information |
| dohanews.co | Yes | Reputed | culture and more in and around Qatar …The documentary features human interest pieces that incorporate the day-to-day lives of Qatar residents |
| iloveqatar.net | Yes | Forum | Qatar National Day - Short Info …the date of December 18 is celebrated each year as the National Day of Qatar… |
| cnn.com | No | Reputed | The 2022 World Cup final in Qatar will be held on December 18 …Qatar will be held on December 18—the Gulf state's national day. Confirm. U.S… |
| icassociat | No | Other | In partnership with ProEvent Qatar, ICA can confirm that the World Stars |
| ion.co.uk | | | will be led on the 17 December, World Stars vs. Qatar Stars - Qatar National Day. |

trained for the task as part of a deep neural network (NN). We also used a task-specific embedding of the question and of the answer together with all the above evidence about it, which comes from the last hidden layer of the neural network.

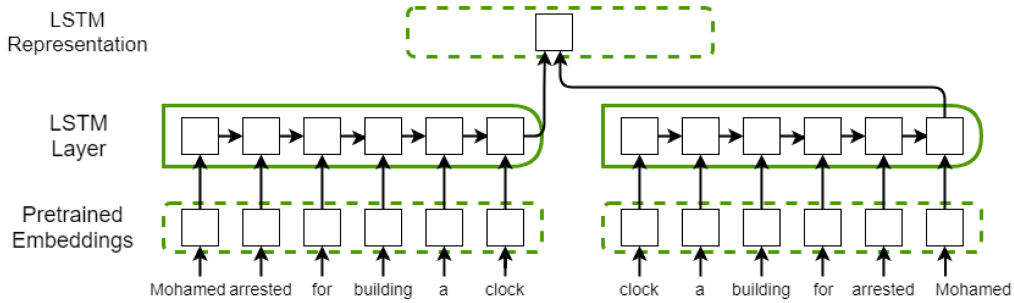
### 3.4 Classification Model

Our model combines an LSTM-based neural network with kernel-based support vector machines. In particular, we use a bi-LSTM recurrent neural network to train abstract feature representations of the examples. We then feed these representations into a kernel-based SVM, together with other features. The architecture is shown in Figure 3(a). We have five LSTM sub-networks, one for each of the text sources from two search engines: *Claim*, *Google Web page*, *Google snippet*, *Bing Web page*, and *Bing snippet*. We feed the claim (i.e., the answer) into the neural network as-is. As we can have multiple snippets, we only use the best-matching one as described above. Similarly, we only use a single best-matching triple of consecutive sentences from a Web page. We further feed the neural network with the similarity features described above. All these vectors are concatenated and fully connected to a much more compact hidden layer that captures the task-specific embeddings. This layer is connected to a softmax output unit that classifies the claim as true or false. Figure 3(b) shows the general architecture of each of the LSTM components. The input text is transformed into a sequence of word embeddings, which are then passed to the bidirectional LSTM layer to obtain a representation for the input text.

Next, we extract the last three layers of the neural network—(*i*) the concatenation layer, (*ii*) the embedding layer, and (*iii*) the classification node—and we feed them into an SVM with a radial basis function kernel (RBF). In this way, we use the neural network to train task-specific embeddings of the input text fragments, and also of the entire input example. Ultimately, this yields a combination of deep learning and task-specific embeddings with RBF kernels.




(a) General overview of the architecture.




(b) Detailed LSTM representation.

Fig. 3. Our NN architecture for fact-checking in cQA. Each green box in 3(a) consists of the bi-LSTM structure in 3(b).

Table 11. Baseline vs. Our Pilot SVM, Predicting One of Three Classes of the Questions (Factual vs. Opinion vs. Just Socializing)

| System | Accuracy |
|---|---|
| Baseline: All OPINION (majority class) | 50.7 |
| Our pilot: SVM, bag-of-words | 62.0 |
| Our pilot: SVM, text features | 60.3 |

## 3.5 Experiments

*3.5.1 Question Classification.* Table 11 shows the results of our run for classification of the three question categories (FACTUAL, OPINION, SOCIALIZING), using an SVM with bag-of-words and some other features. We can see a 10-point absolute improvement over the baseline, which means the task is feasible. This also leaves plenty of space for further improvement, which is beyond the scope of this work. Instead, below we focus on the more interesting task of checking the factuality of *Good* answers to *Factual* questions.

*3.5.2 Answer Classification.*

*Setting and Evaluation.* We perform leave-one-thread-out cross-validation, where each time we exclude and use for testing one of the 71 questions together with all its answers. This is done to

Table 12. Fact-checking the Answers in a cQA Forum, Focusing on the Impact of the Contextual and Discourse Features

| Our System | Accuracy | Precision | Recall | $F_1$ |
|---|---|---|---|---|
| All information sources | 0.683 | 0.693 | 0.688 | 0.690 |
| All\discourse | 0.659 | 0.675 | 0.648 | 0.661 |
| All\context | 0.574 | 0.583 | 0.602 | 0.592 |
| All\discourse and context | 0.542 | 0.554 | 0.563 | 0.558 |
| Only context+discourse | 0.635 | 0.621 | 0.742 | 0.676 |
| **Baseline** | | | | |
| *All positive (majority class)* | 0.514 | 0.514 | 1.000 | 0.679 |

respect the structure of the threads when splitting the data. We report Accuracy, Precision, Recall, and $F_1$ for the classification setting.

We used a bidirectional LSTM with 25 units and a hard-sigmoid activation, which we trained using an RMSprop optimizer with 0.001 initial learning rate, L2 regularization with $\lambda = 0.1$, and 0.5 dropout after the LSTM layers. The size of the hidden layer was 60 with *tanh* activations. We used a batch of 32 and we trained for 400 epochs. Similarly to the bi-LSTM layers, we used an $l_2$ regularizer with $\lambda = 0.01$ and dropout with a probability of 0.3.

For the SVM, we used grid search to find the best parameters for the parameters $c$ and $\gamma$. We optimized the SVM for classification Accuracy.

*Results.* Table 12 shows the results from our experiments for several feature combinations and for two baselines. First, we can see that our system with all features performs better than the baseline for Accuracy. The ablation study shows the importance of the context and of the discourse features. When we exclude the discourse and the contextual features, the accuracy drops from 0.683 to 0.659 and 0.574, respectively. When both the context and the discourse features are excluded, the accuracy drops even further, to 0.542. The $F_1$ results are consistent with this trend. This is similar to the trend for check-worthiness estimation (cf. Table 4). Finally, using the discourse and the contextual features, without any other features, yields an accuracy of 0.635, which is quite competitive. Overall, these results show the importance of the contextual and of the discourse features for the fact-checking task, with the former being more important than the latter.

## 4 DISCUSSION

Here, we look at some examples that illustrate how context and discourse help for our two tasks.

### 4.1 Impact of Context

First, we give some examples where the use of contextual information yields the correct prediction for the check-worthiness task (Section 2). In each of these examples, there is a particular contextual feature type that turned out to be critical for making the correct prediction, namely that these are check-worthy sentences (they were all misclassified as not check-worthy when excluding that feature type):

***Metadata—using opponent's name.*** According to our explanation in Section 2.2, the target sentence in Figure 4(a) mentions the name of the opponent, and this turned out to be the critical feature for correctly predicting that these are check-worthy claims.

***Contradiction.*** Sometimes, an important claim contains a contradiction to what has been said earlier, e.g., the bold sentence in Figure 4(b). We model the contradiction as explained in Section 2.2 to extract such check-worthy claims.

(a) Clinton: They're doing it to try to influence the election for **Donald Trump**.

(b) Clinton: In the days after the first debate, you sent out a series of tweets from 3 a.m. to 5 a.m., including one that told people to check out a sex tape.
Clinton: Is that the discipline of a good leader?
Trump **No, there wasn't check out a sex tape.**

Fig. 4. Examples of the impact of context in debate instances.

(a) $q$: *what is qtel customer service number?*
$a_1$: Call 111 … and hold the line for at least 30 minutes before being answered. Enjoy Q-tel music.
$a_2$: call 111
$a_3$: 111 - Customer service
$a_4$: 111

(b) $q$: I have visiting visa for work; so can I drive? I have egyptian license
$a$: If you are on a visiting Visa and and you have an international driver license you can use it for 6 month I guess.

**Evidence**: *[…] A valid international driving license can be used from the day you arrive into the country until six months. […]*

(c) $q$: Smoke after dark during ramadan?
$a$: Yes! You can smoke in public after sunset till dawn.

**Evidence**: *Bars and pubs will generally remain open but will only serve alcohol after dark. […] Don't smoke, drink, chew gum or eat in public during the hours of sunrise to sunset.*

(d) $q$: I am in the process of coming over to work in Doha. I wanted to know if their were any specific shots and vaccinations I should get before coming over. I want to try and get these things completed before I leave the US.
$a$: Yes there are; though it varies depending on which country you come from. In the UK; the doctor has a list of all countries and the vaccinations needed for each. I'll imagine they have the same in the US.

Fig. 5. Examples of the impact of context and discourse in cQA instances.

***Similarity of the sentence to known positive/negative examples.*** The sentence *"For the last seven-and-a-half years, we've seen America's place in the world weakened"* is similar to the already fact-checked sentence *"We've weakened America's place in the world."* Thus, the latter is to be classified as check-worthy.

Following, there are some examples for the cQA fact-checking task, where the use of particular contextual features allowed the system to predict correctly the factuality of the answers (they were all misclassified when the corresponding contextual feature was turned off):

***Support from the current thread.*** The example in Figure 5(a) shows how the thread information (e.g., similarity of one answer to the other answers in the thread) helps to predict the answer's factuality. The question has four answers that should all be True, but they had been misclassified without the support from the current thread.

***Support from high-quality posts in Qatar Living.*** The example in Figure 5(b) was correctly classified as True when using the high-quality posts and misclassified as False otherwise. The high-quality posts in the QL forum contain verified information about common topics discussed by people living in Qatar, such as visas, driving regulations, customs, and so on. The example shows one piece of relevant evidence selected by our method from the high-quality posts, which possibly helps in making the right classification.

***Support from all of Qatar Living***. The example in Figure 5(c) shows the evidence found in the search results in the entire Qatar Living forum. It was classified correctly as True when using the support from all of the Qatar Living forum, and it was misclassified without it.

## 4.2 Impact of Discourse

As the evaluation results have shown, discourse also played an important role. Let us take the check-worthiness task as an example. In the sentence *"But what President Clinton did, he was impeached, he lost his license to practice law,"* the discourse parser identified the fragment "*But what President Clinton did*" as Background referring to the text for facilitating understanding; the segment "*he was impeached*" is Elaboration referring to additional information, and "*. . . to practice law*" is Enablement referring to the action. These relations are associated with factually true claims.

Similarly, for cQA fact-checking, using discourse information yielded correct classification as TRUE for the example in Figure 5(d). The question and the answer were parsed together, and the segment containing the answer was identified as ELABORATION. The answer further contains a BACKGROUND segment ("*In the UK, the doctor has a list of all countries and the vaccinations needed for each.*") and an ATTRIBUTION segment ("*they have the same in the US*"). These discourse relations are also associated with factually true answers (as we have seen also in Figure 5(c)).

## 5 RELATED WORK

Journalists, Web users, and researchers are aware of the proliferation of false information on the Web, and as a result, topics such as information credibility and fact-checking are becoming increasingly important as research directions (Lazer et al. 2018; Vosoughi et al. 2018). For instance, there was a recent special issue of the *ACM Transactions on Information Systems* journal on Trust and Veracity of Information in Social Media (Papadopoulos et al. 2016), there was a SemEval-2017 shared task on Rumor Detection (Derczynski et al. 2017), and there was a lab at CLEF-2018 on Automatic Identification and Verification of Claims in Political Debates (Atanasova et al. 2018; Barrón-Cedeño et al. 2018; Nakov et al. 2018).

### 5.1 Detecting Check-Worthy Claims

The task of detecting check-worthy claims has received relatively little research attention so far. Hassan et al. (2015) developed *ClaimBuster*, which assigns each sentence in a document a score, i.e., a number between 0 and 1 showing how worthy it is for fact-checking. The system is trained on their own dataset of about 8,000 debate sentences (1,673 of them check-worthy), annotated by students, university professors, and journalists. Unfortunately, this dataset is not publicly available, and it contains sentences without context, as about 60% of the original sentences had to be thrown away due to lack of agreement. In contrast, we developed a new publicly available dataset based on manual annotations of political debates by nine highly reputed fact-checking sources, where sentences are annotated in the context of the entire debate. This allows us to explore a novel approach, which focuses on the context. Note also that the *ClaimBuster* dataset is annotated following guidelines from Hassan et al. (2015) rather than trying to mimic a real fact-checking website; yet, it was later evaluated against PolitiFact (Hassan et al. 2016). In contrast, we train and evaluate directly on annotations from fact-checking websites, and thus we learn to fit them better.[22]

Patwari et al. (2017) also focused on the 2016 US Presidential election campaign and independently obtained their data following a similar approach. Their setup asked to predict whether any of the fact-checking sources 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 was that coming from the most confident classifier. The features considered go from LDA topic-modeling to POS tuples and bag-of-word representations. Unlike that work, we further mimic the selection strategy of one particular fact-checking organization by learning to jointly predict the selection choices by various such organizations.

The above-mentioned lab on fact-checking at CLEF-2018 was partially based on a variant of our data, but it focused on one fact-checking organization only (Atanasova et al. 2018), unlike our multi-source setup here.

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

[22]Our model is released as an online demo that supports both English and Arabic (Jaradat et al. 2018): http://claimrank.qcri.org/.

of a statement and what is believed, e.g., "falsely claimed that X." The claims matching the patterns go through a statistical classifier that marks the text of the claim. This procedure can be used to acquire a corpus of disputed claims from the Web. Given a set of disputed claims, Ennals et al. (2010b) approached the task as locating new claims on the Web that entail the ones that have already been collected. Thus, the task can be conformed as recognizing textual entailment, which is analyzed in detail in Dagan et al. (2009). Finally, Le et al. (2016) argued that the top terms in claim vs. non-claim sentences are highly overlapping, which is a problem for bag-of-words approaches. Thus, they used a CNN, where each word is represented by its embedding and each named entity is replaced by its tag, e.g., *person*, *organization*, *location*.

### 5.2 Fact-Checking and Credibility

The credibility of contents on the Web has been questioned by researchers for a long time. While in the early days the main research focus was on online news portals (Brill 2001; Hardalov et al. 2016), the interest has eventually shifted towards social media (Castillo et al. 2011; Karadzhov et al. 2017a; Popat et al. 2017; Vosoughi et al. 2018; Zubiaga et al. 2016), which are abundant in sophisticated malicious users such as opinion manipulation *trolls* (Mihaylov et al. 2018)—paid (Mihaylov et al. 2015b) or just perceived (Mihaylov et al. 2015a; Mihaylov and Nakov 2016b)—*sockpuppets* (Maity et al. 2017), *Internet water army* (Chen et al. 2013), and *seminar users* (Darwish et al. 2017).

Most of the efforts on assessing credibility have focused on micro-blogging websites. For instance, Canini et al. (2011) studied the credibility of Twitter accounts (as opposed to tweet posts), and found that both the topical content of information sources and social network structure affect source credibility. Another work, closer to ours, aims at addressing credibility assessment of rumors on Twitter as a problem in finding false information about a newsworthy event (Castillo et al. 2011). Their model considered a variety of features including user reputation, writing style, and various time-based features, among others.

Other efforts have focused on news communities. For example, several truth-discovery algorithms were studied and combined in an ensemble method for veracity estimation in the VERA system (Ba et al. 2016). They proposed a platform for end-to-end truth discovery from the Web: extracting unstructured information from multiple sources, combining information about single claims, running an ensemble of algorithms, and visualizing and explaining the results. They also explored two different real-world application scenarios for their system: fact-checking for crisis situations and evaluation of trustworthiness of a rumor. However, the input to their model is structured data, while here we are interested in unstructured text. Similarly, the task defined in Mukherjee and Weikum (2015) combines three objectives: assessing the credibility of a set of posted articles, estimating the trustworthiness of sources, and predicting users' expertise. They considered a manifold of features characterizing language, topics, and Web-specific statistics (e.g., review ratings) on top of a continuous conditional random fields model. In follow-up work, Popat et al. (2016) proposed a model to support or refute claims from Snopes.com and Wikipedia by considering supporting information gathered from the Web. In another follow-up work, Popat et al. (2017) proposed a complex model that considers stance, source reliability, language style, and temporal information.

Another important research direction is on using tweets and temporal information for checking the factuality of rumors. For example, Ma et al. (2015) used temporal patterns of rumor dynamics to detect false rumors and to predict their frequency. They focused on detecting false rumors in Twitter using time series. They used the change of social context features over a rumor's life cycle to detect rumors at an early stage after they were broadcast.

A more general approach for detecting rumors is explored in Ma et al. (2016), who used recurrent neural networks to learn hidden representations that capture the variation of contextual

information of relevant posts over time. Unlike this work, we do not use microblogs, but we query the Web directly in search for evidence.

In the context of question answering, there has been work on assessing the credibility of an answer, e.g., based on intrinsic information, i.e. without any external resources (Banerjee and Han 2009). In this case, the reliability of an answer is measured by computing the divergence between language models of the question and of the answer. The spawn of community-based question-answering websites also allowed for the use of other kinds of information. Click counts, link analysis (e.g., PageRank), and user votes have been used to assess the quality of a posted answer (Agichtein et al. 2008; Jeon et al. 2006; Jurczyk and Agichtein 2007). Nevertheless, these studies address the answers' credibility level just marginally.

Efforts to estimate the credibility of an answer to assess its overall quality required the inclusion of content-based information (Su et al. 2010), e.g., verbs and adjectives such as *suppose* and *probably*, which cast doubt on the answer. Similarly, Lita et al. (2005) used source credibility (e.g., does the document come from a government website?), sentiment analysis, and answer contradiction compared to other related answers. Another way to assess the credibility of an answer is to incorporate textual-entailment methods to find out whether a text (question) can be derived from a hypothesis (answer). Overall, the *credibility* assessment for question answering has been mostly modeled at the feature level, with the goal of assessing the quality of the answers. A notable exception is the work of Nakov et al. (2017b), where credibility is treated as a task in its own right. Yet, *credibility* is different from *factuality* (our focus here), as the former is a subjective perception about whether a statement is credible, rather than verifying it as true or false; still, these notions are often wrongly mixed in the literature. To the best of our knowledge, no previous work has targeted fact-checking of answers in the context of community Question Answering by gathering external support.

## 6 CONCLUSION AND FUTURE WORK

We have studied the role of context and discourse information for two factuality tasks: (*i*) detecting check-worthy claims in political debates, and (*ii*) fact-checking answers in a community question-answering forum. We have developed annotated resources for both tasks, which we have made publicly available, and we have proposed rich input representations—including discourse and contextual features—and also a complementary set of core features to make our systems as strong as possible. The definition of context varies between the two tasks. For check-worthiness estimation, a target sentence occurs in the context of a political debate, where we model the current intervention by a debate participant in relation to the previous and to the following participants' turns, together with meta information about the participants, about the reaction of the debate's public, and so on. In the answer's factuality checking task, the context for the answer involves the full question-answering thread, the related threads in the entire forum, or the set of related high-quality posts in the forum.

We trained classifiers for both tasks using neural networks, kernel-based support vector machines, and combinations thereof, and we ran a rigorous evaluation, comparing against alternative systems whenever possible. We also discussed several cases from the test set where the contextual information helped make the right decisions. Overall, our experimental results and the posterior manual analysis have shown that discourse cues, and especially modeling the context, play an important role and thus should be taken into account when developing models for these tasks.

In future work, we plan to study the role of context and discourse for other related tasks, e.g., for checking the factuality of general claims (not just answers to questions), and for stance classification in the context of factuality. We also plan to experiment with a joint model for

check-worthiness estimation, for stance classification, and for fact-checking, which would be useful in an end-to-end system (Baly et al. 2018; Mohtarami et al. 2018).

We would also like to extend our datasets (e.g., with additional debates, but also with interviews and general discussions), thus enabling better exploitation of deep learning. Especially for the answer-verification task, we would like to try distant supervision based on known facts, e.g., from high-quality posts, which would allow us to use more training data. We also want to improve user modeling, e.g., by predicting factuality for the user's answers and then building a user profile based on that. Finally, we want to explore the possibility of providing justifications for the verified answers, and ultimately of integrating our system in a real-world application.

## REFERENCES


Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the International Conference on Web Search and Data Mining (WSDM'08)*. 183–194.

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*.

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 (WWW'16)*. 159–162.

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 (NAACL-HLT'18)*. 21–27.

Protima Banerjee and Hyoil Han. 2009. Answer credibility: A language modeling approach to answer validation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'09)*. 157–160.

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*.

Jonathan Baxter. 1997. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Mach. Learn.* 28, 1 (1997), 7–39.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3, 1 (2003), 993–1022.

Ann M. Brill. 2001. Online journalists embrace new marketing function. *News. Res. J.* 22, 2 (2001), 28.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: Experiences from the scikit-learn project. In *Proceedings of the ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 108–122.

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 (SocialCom/PASSAT'11)*. 1–8.

Richard Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the 10th International Conference on Machine Learning (ICML'13)*. 41–48.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW'11)*. 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 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'13)*. 116–120.

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

Kareem Darwish, Dimitar Alexandrov, Preslav Nakov, and Yelena Mejova. 2017. Seminar users in the Arabic Twitter Sphere. In *Proceedings of the 9th International Conference on Social Informatics (SocInfo'17)*. 91–108.

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 (SemEval'17)*. 60–67.

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 (WICOW'10)*. 67–74.

Rob Ennals, Beth Trushkowsky, and John Mark Agosta. 2010b. Highlighting disputed claims on the web. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. 341–350.

George Foster and Roland Kuhn. 2009. Stabilizing minimum error rate training. In *Proceedings of the 4th Workshop on Statistical Machine Translation (StatMT'09)*. 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 Recent Advances in Natural Language Processing (RANLP'17)*. 267–276.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (PMLR'11)*, Vol. 15. 315–323.

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 (AIMSA'16)*. 172–180.

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

Naeemul Hassan, Mark Tremayne, Fatma Arslan, and Chengkai Li. 2016. Comparing automated factual claim detection against judgments of journalism organizations. In *Proceedings of the Computation + Journalism Symposium*.

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

Ken Hyland. 2005. *Metadiscourse: Exploring Interaction in Writing*. Bloomsbury Publishing.

Daisuke Ishikawa, Tetsuya Sakai, and Noriko Kando. 2010. Overview of the NTCIR-8 community QA pilot task (Part I): The test collection and the task. In *Proceedings of NTCIR-8 Workshop Meeting*. 421–432.

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: Human Language Technologies (NAACL-HLT'18)*. 26–30.

Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee, and Soyeon Park. 2006. A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*. 228–235.

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

Pawel Jurczyk and Eugene Agichtein. 2007. Discovering authorities in question answer communities by using link analysis. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM'07)*. 919–922.

Georgi Karadzhov, Pepa Gencheva, Preslav Nakov, and Ivan Koychev. 2017a. We built a fake news & click-bait filter: What happened next will blow your mind!. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'17)*. 334–343.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017b. Fully automated fact checking using external sources. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'17)*. 344–353.

Lauri Karttunen. 1971. Implicative verbs. *Language* 47, 2 (1971), 340–358.

Milen Kouylekov and Matteo Negri. 2010. An open-source package for recognizing textual entailment. In *Proceedings of the 48th Meeting of the Association for Computational Linguistics: System Demonstrations (ACL'10)*. 42–47.

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 (2018), 1094–1096.

Dieu-Thu Le, Ngoc Thang Vu, and Andre Blessing. 2016. Towards a text analysis system for political debates. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH'16)*. 134–139.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

Lucian Vlad Lita, Andrew Hazen Schlaikjer, WeiChang Hong, and Eric Nyberg. 2005. Qualitative dimensions in question answering: Extending the definitional QA task. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI'05)*. 1616–1617.

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 (WWW'05)*. 342–351.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics (ETMTNLP'02)*. 63–70.

Caroline Lyon, James Malcolm, and Bob Dickerson. 2001. Detecting short passages of similar text in large document collections. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'01)*.

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 International Joint Conference on Artificial Intelligence (IJCAI'16)*. 3818–3824.

Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM'15)*. 1751–1754.

Suman Kalyan Maity, Aishik Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2017. Detection of sockpuppets in social media. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'17)*. 243–246.

Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. 2015a. Finding opinion manipulation trolls in news community forums. In *Proceedings of the 19th Conference on Computational Natural Language Learning (CoNLL'15)*. 310–314.

Todor Mihaylov, Ivan Koychev, Georgi Georgiev, and Preslav Nakov. 2015b. Exposing paid opinion manipulation trolls. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'15)*. 443–450.

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

Todor Mihaylov and Preslav Nakov. 2016b. Hunting for troll comments in news community forums. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics (ACL'16)*. 399–405.

Todor Mihaylov and Preslav Nakov. 2016a. SemanticZ at SemEval-2016 Task 3: Ranking relevant answers in community question answering using semantic similarity based on fine-tuned word embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 879–886.

Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadjov, and James Glass. 2018. Fact checking in community forums. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*. 5309–5316.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *Retrieved from CoRR* abs/1309.4168.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'13)*. 746–751.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8640.2012.00460.x.

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 16th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'18)*. 767–776.

Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM'15)*. 353–362.

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 the 9th International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction (Lecture Notes in Computer Science)*. 372–387.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017a. SemEval-2017 Task 3: Community question answering. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval'17)*. 27–48.

Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 2015. SemEval-2015 Task 3: Answer selection in community question answering. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval'15)*. 269–281.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 Task 3: Community question answering. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval'16)*. 525–545.

Preslav Nakov, Tsvetomila Mihaylova, Lluís Màrquez, Yashkumar Shiroya, and Ivan Koychev. 2017b. Do not trust the trolls: Predicting credibility in community question answering forums. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'17)*. 551–560.

Symeon Papadopoulos, Kalina Bontcheva, Eva Jaho, Mihai Lupu, and Carlos Castillo. 2016. Overview of the special issue on trust and veracity of information in social media. *ACM Trans. Inf. Syst.* 34, 3 (2016), 14:1–14:5.

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 ACM Conference on Information and Knowledge Management (CIKM)*. 2259–2262.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM'16)*. 2173–2178.

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 (WWW'17)*. 1003–1012.

Martin Potthast, Matthias Hagen, Tim Gollub, Martin Tippmann, Johannes Kiesel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2013. Overview of the 5th international competition on plagiarism detection. In *Proceedings of the CLEF Conference on Multilingual and Multimodal Information Access Evaluation*. 301–331.

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 Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*. 2931–2937.

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

Radim Řehuřek and Petr Sojka. 2010. Software framework for topic modelling with large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. 45–50.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03)*. 105–112.

Qi Su, Helen Kai-Yun Chen, and Chu-Ren Huang. 2010. Incorporate credibility into context for the best social media answers. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC'10)*. 535–541.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML'13)*, Vol. 28. 1139–1147.

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

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 (03 2016), 1–29.

Received May 2018; revised October 2018; accepted November 2018