Noname manuscript No.
(will be inserted by the editor)

Exploiting stance hierarchies for cost-sensitive stance detection of Web documents

Arjun Roy · Pavlos Fafalios · Asif Ekbal · Xiaofei Zhu · Stefan Dietze

Received: date / Accepted: date

Abstract Fact checking is an essential challenge when combating fake news. Identifying documents that agree or disagree with a particular statement (claim) is a core task in this process. In this context, stance detection aims at identifying the position (stance) of a document towards a claim. Most approaches address this task through classification models that do not consider the highly imbalanced class distribution. Therefore, they are particularly ineffective in detecting the minority classes (for instance, 'disagree'), even though such instances are crucial for tasks such as fact-checking by providing evidence for detecting false claims. In this paper, we exploit the hierarchical nature of stance classes which allows us to propose a modular pipeline of cascading binary classifiers, enabling performance tuning on a per step and class basis. We implement our approach through a combination of neural and traditional classification models that highlight the misclassification costs of minority classes. Evaluation results demonstrate state-of-the-art performance of our approach and its ability to significantly improve the classification performance of the important 'disagree' class.

Keywords Stance detection · Fact-checking · Cascading classifiers · Fake News

A. Roy
L3S Research Center, Leibniz University of Hannover, Hannover, Germany
E-mail: roy@L3S.de

P. Fafalios
Institute of Computer Science, FORTH-ICS, Heraklion, Greece
E-mail: fafalios@ics.forth.gr

A. Ekbal
Dept. of CSE, IIT Patna, Patna, India
E-mail: asif@iitp.ac.in

X. Zhu
Chongqing University of Technology, Chongqing, China
E-mail: zxf@cqut.edu.cn

S. Dietze
GESIS - Leibniz Institute for the Social Sciences, Cologne & Heinrich-Heine-University Düsseldorf, Germany
E-mail: stefan.dietze@GESIS.org

1 Introduction

Spread of fake news and false claims have become ubiquitous due to the widespread use and network effects facilitated by social online platforms [25]. A recent study has shown that false claims are re-tweeted faster, further, and for longer than true claims on twitter [42]. Another study found that the top 20 fake news stories about the 2016 U.S. presidential election received more engagement on Facebook than the top 20 election stories from the 19 major media outlets [8]. These findings demonstrate the significance and scale of the fake news problem and the potential effects it can have on contemporary society [2].

In this context, *fact-checking* is the task of determining the veracity of an assertion or statement (a *claim*) [34,45,21]. Identifying the particular semantics of relation between a (Web) document and a given statement is an essential task, which is widely referred to as *stance detection* [23]. More precisely, stance detection aims at identifying the stance (perspective) of a document towards a claim, namely whether the document *agrees* with the claim, *disagrees* with the claim, *discusses* about the claim without taking a stance, or is entirely *unrelated* to the claim. However, the real-world distribution of the aforementioned classes is highly imbalanced, where in particular instances of classes of crucial importance (*related*, *disagree*) are strongly underrepresented. This is reflected by state-of-the-art benchmark datasets such as the *FNC-I* dataset [33], where the *disagree* class corresponds to less than 3% of the instances.

Existing approaches that try to cope with this problem are ineffective in detecting instances of minority classes. For instance, whereas the overall performance of state-of-the-art systems [4,17,35,6,18,47,27] ranges between 58% and 61% (F1 macro-average), the performance on the *disagree* class ranges between 3% and 18% only. However, this class is of key importance in fact-checking since it enables detecting documents that provide evidence for invalidating false claims. It is worth noting that there also have been efforts to deal with a similar problem by simply introducing a class-wise penalty into the loss function [26,41] , thereby limiting its capacity to solve the class imbalance problem.

In this paper, we exploit the hierarchical nature of stance classes (Figure 1) by introducing a classifier cascade of three binary classification models, where individual tuning at each step enables better consideration of misclassification costs of minority classes. This is aimed at improving the classification performance of the important but underrepresented *disagree* class without negatively affecting the performance of other classes. To this end, we propose a three-stage pipeline architecture that treats the 4-class classification problem as three different binary classification (sub-)tasks of increasing difficulty. The first stage aims at detecting the documents related to the claim (*relevance classification*), the second stage classifies only related documents and focuses on detecting those that take a stance towards the claim (*neutral/stance classification*), and the third stage classifies documents taking a stance as agreeing or disagreeing to the claim (*agree/disagree classification*). This modular and step-wise approach offers the flexibility to consider different classifiers and features in each different stage of the pipeline depending on the context and corpus at hand, thus enabling to optimise performance on a per stage and class basis.

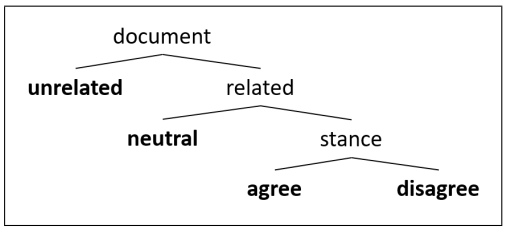

Fig. 1 Document stance hierarchy.

We provide an implementation[1] of our pipeline architecture, called *L3S (Learning in 3 Steps)*, using a combination of neural and traditional classification models and by introducing cost-sensitive measures to reflect the importance of minority classes. Evaluation results following the established stance detection benchmark[33] show that our approach achieves the state-of-the-art performance on the general stance classification problem, slightly outperforming all the existing methods by one percentage point (macro-average F1 score). Most importantly, we significantly improve the F1 score of the *disagree* class by 28% compared to the state-of-the-art.

The rest of the paper is organized as follows: Section 2 formulates the problem and provides an overview of our pipeline approach. Section 3 describes supervised models for each stage of the pipeline. Section 4 reports evaluation results. Section 5 presents related works. Finally, Section 6 concludes the paper.

2 Problem Description and Approach Overview

Given a textual claim c (e.g., *"KFC restaurants in Colorado will start selling marijuana"*) and a document d (e.g., an article), stance classification aims at classifying the stance of d towards c to one of the following four categories (classes):

- **Unrelated**: the document is not related to the claim.
- **Neutral**: the document discusses about the claim but it does not take a stance towards its validity.
- **Agree**: the document agrees with the claim.
- **Disagree**: the document disagrees with the claim.

These four classes can be structured in a tree-like hierarchy as shown in Fig. 1. At first, a document can be either *unrelated* or *related* to the claim. Then, a document that is related to the claim can either be *neutral* to the claim or take a *stance*. Finally, a document that takes a stance can either *agree* or *disagree* with the claim. The leaves of the tree are the four classes. Considering this structure, we can now model the stance classification problem as a three-stage classifier cascade consisting of three connected sub-tasks (or *stages*):

- **Stage 1 (relevance classification)**: identify if a document is related to the claim or not.
- **Stage 2 (neutral/stance classification)**: identify if a document classified as *related* from stage 1 is neutral to the claim or takes a stance.

[1] `https://github.com/arjunroyihrpa/Stance-Hierarchies`

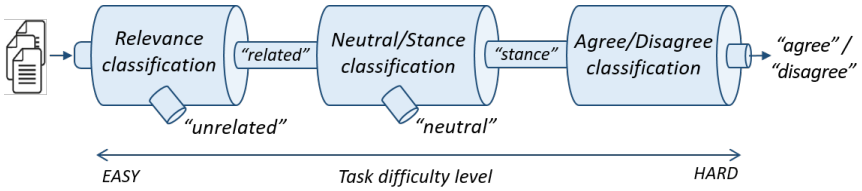

Fig. 2 Pipeline for document stance classification and difficulty level of each stage.

– **Stage 3 (agree/disagree classification)**: identify if a document classified as *stance* from stage 2 agrees or disagrees with the claim.

Figure 2 depicts the proposed pipeline architecture. Each stage of the pipeline can now be modeled as a separate binary classification problem. We hypothesize that relevance classification (stage 1) can filter out documents unrelated to the claim which can facilitate the neutral/stance classification task (stage 2). Likewise, knowing that a document takes a stance towards the claim can facilitate the agree/disagree classification task (stage 3). A weakness of such a pipeline approach is that errors can propagate from one stage to the other, thus errors in earlier stages negatively affect the later stages. It is to be noted though that, in general, relevance classification (stage 1) is a much easier task than neutral/stance classification (stage 2), which in turn is considered easier than agree/disagree classification (stage 3). Our experimental evaluation validates this hypothesis (more in Sect. 4).

3 Pipeline Implementation

In this section we describe our classification models for each stage of the pipeline, each being implemented through a supervised model tailored towards the specific classification problem at hand and using features that reflect the syntactical and semantic similarity between a claim and a document.

3.1 Stage 1: Relevance Classification

Existing approaches on separating the *related* from the *unrelated* instances have already achieved a high accuracy (>95%) [18]. Such models usually make use of hand-crafted features that aim at reflecting the text similarity between the claim and the document. Since our focus is on the important *disagree* class (which is part of *related* in this stage), we seek a model that penalises misclassifications of this class. To this end, we experimented with a variety of different classifiers, inspired by previous works that perform well on the same problem, including: support vector machine (SVM) with and without class-wise penalty, gradient boosting trees, AdaBoost, decision trees, random forest with and without class-wise penalty, and convolutional neural networks (CNN).

To train the classifiers, we use the below set of features, selected through extensive feature analysis. The first four features are used in the baseline model provided by FNC-I [33], the next two are inspired from [44], and the last two (*keyword*, *proper noun overlap*) are new.

- *N-grams match:* A *n-gram* is the sequence of n continuous words in a given text. The feature value is defined as the number of common n-grams in the claim and the document. For our system, we choose bigrams, trigrams and fourgrams.
- *Chargrams match:* Similar to n-gram, chargram is a sequence of n continuous characters. We use bi-, tri- and four-chargrams in our system.
- *Binary co-occurrence:* This feature consists of two values. The first one is the number of words of the claim that appear in the first 255 words of the document, and the second one is the number of words in the claim that appear in the entire body of the document.
- *Lemma overlap:* This feature is similar to the unigram match with the difference that the words are first converted into their lemmatized form.
- *Text similarity:* We calculate the cosine similarity between the text of the claim and each sentence of the document. The maximum similarity value is considered as the feature value.
- *Word2vec similarity:* The cosine similarity between the pre-trained word2vec embeddings [28] of the claim and the document (reflecting their semantic relationship).
- *Keyword overlap:* We extract important words that appear in the text of the claim and the text of the document using *cortical.io*[2]. The feature is defined as the number of common keywords in the claim and the document.
- *Proper noun overlap:* This feature is same as keyword overlap but instead of keywords we extract proper nouns using the NLTK Part-of-Speech tagger.[3]

Through cross validation, we found that a simple SVM classifier with class-wise penalty [41] outperforms all the other models. In more detail, we solve the following optimization problem:

$$Min_{\varpi,\beta}(\frac{\varpi^T\varpi}{2} + \alpha_1 \sum_{i=1|\gamma_i \in Relat.}^{m} \epsilon_i + \alpha_2 \sum_{i=1|\gamma_i \in Unrel.}^{m} \epsilon_i) \quad (1)$$

Subjected to the constraints:

$$\gamma_i(\varpi^T\chi_i + \beta) \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0, \quad i = 1, ..., m \quad (2)$$

where, ϖ is weight vector, β is bias, γ_i is the output constraint function, χ_i is the training input vector, and α_1 and α_2 are regularization hyperparameters of penalty terms ϵ_i for *related* and *unrelated* class, respectively. A pictorial representation of the model is depicted in Fig. 3. Hyperparameters are tuned through 10-fold cross-validation on the training dataset.

3.2 Stage 2: Neutral/Stance Classification

In this stage, we require a model that is able to consider the misclassification costs of minority classes, which is vitally important for fact-checking tasks. Previous works on similar problems have shown that deep learning models [36,3], supervised classifiers with sentiment-related features [24], as well as sentiment features

[2] `https://www.cortical.io/`
[3] `https://www.nltk.org/`

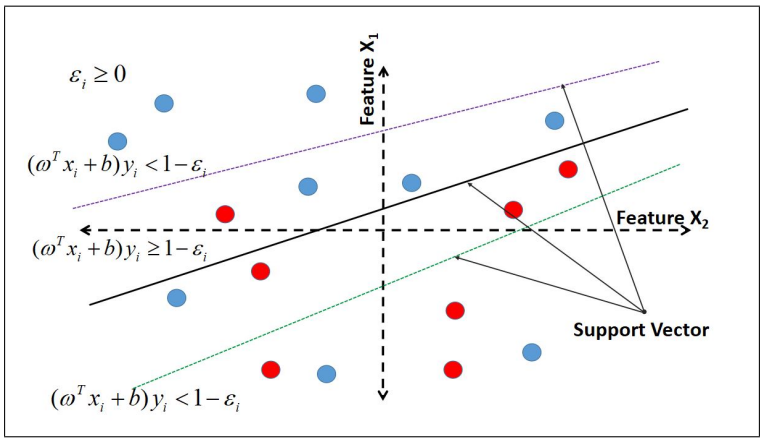

Fig. 3 Diagram outlining the SVM model used in stages 1 and 3.

employed in neural models [40], help solving the problem effectively. We tried a variety of approaches including both neural models (Bi-LSTM, CNN) and classical machine learning models (like SVM, gradient boosting and decision trees). Based on results on a validation set, we found that a simple CNN model with embedded word vectors and sentiment features outperforms all other (and more complex) models. The top performing system of FNC-I also uses a CNN for the overall 4-class stance classification problem.

To generate sentiment scores for the claims and documents, we use the NLTK sentiment intensity analyzer [22]. NLTK applies a rule-based model for sentiment analysis, and is preferred in comparison to the other systems because it is very effective in aggregating the sentiment polarity of multiple negative words. Our intuition is that a document supporting or refuting a claim will have a strong overall sentiment polarity, while a document not taking a stance towards the claim will have a more neutral overall polarity. In addition, by inspecting several documents that take a stance we noticed that the stance is usually expressed in the first few lines (this is also supported in [11]). Thus, we consider only the document's first 10 sentences in our analysis.

For implementing the CNN, we do not follow the approach used by the top performing system of FNC-I because the training data in our case is limited (stage 1 has filtered out *unrelated* instances), and their CNN model underfits and learns only the *neutral* class. Instead, we propose a simple while effective network architecture, and its graphical representation is depicted in Fig. 4. In particular, we first convert the text of the claim and the first part of the document into embedded tensors (C and D, respectively), using the word representation method proposed by [28]. Next, we generate an array of four sentiment scores (positive, negative, neutral, compound) for both the claim and the document using NLTK (arrays S_C and S_D, respectively). The two vectors C and D are passed through two separate convolutional networks $H1_s^{(f)}$ and $H2_z^{(f)}$, where s and z is the stride number in each case, each with f number of $\eta \times d$ filters (where each of f filters strides through

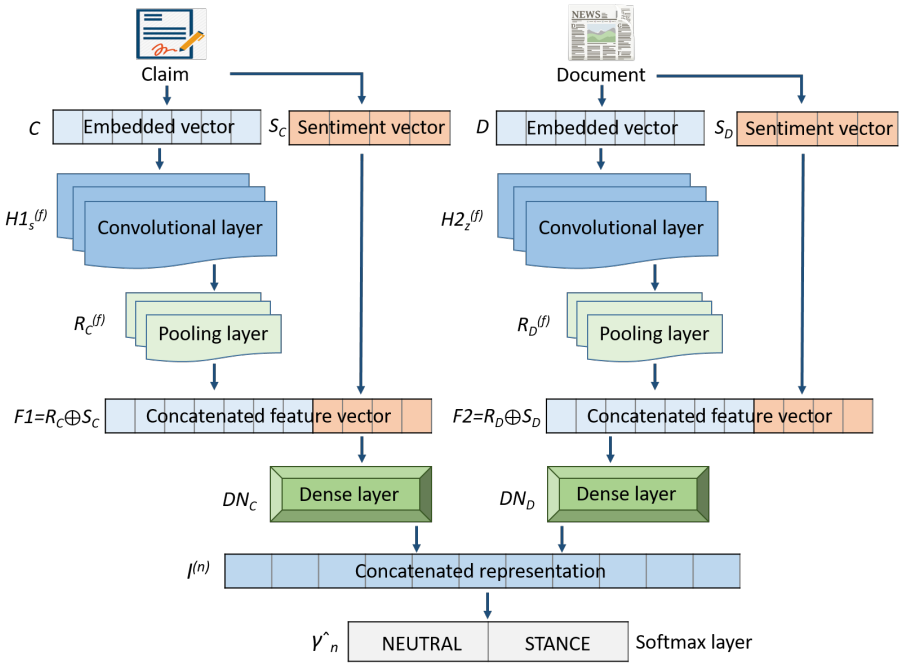

Fig. 4 Diagram outlining the CNN model used in Stage 2.

η words at a time) and stride length of 1, i.e.:

$$H1_s^{(f)} = \phi(\sum_{i=0}^{\eta-1}\sum_{j=0}^{d} \varpi_{i,j}^{(f)} \cdot C_{s+i,j} + \beta^{(f)}) \tag{3}$$

$$H2_z^{(f)} = \phi(\sum_{i=0}^{\eta-1}\sum_{l=0}^{d} \varpi_{i,l}^{(f)} \cdot D_{z+i,l} + \beta^{(f)}) \tag{4}$$

where $s = 1,...,k-\eta+1$ (k =word length of claim); $z = 1,...,p-\eta+1$ (p =word length of document); η is each filter size; d is dimension of embedding; $\varpi_{i,j}^{(f)}$ and $\varpi_{i,l}^{(f)}$ are weights of (neuron in position i,j and i,l respectively) filter f connecting with $C_{s+i,j}$ and $D_{z+i,l}$ respectively; β is bias; f is filter number; and ϕ is a nonlinear activation function.

These are then followed by corresponding global max-pooling layers:

$$R_C^{(f)} = max(H1^{(f)}) \tag{5}$$

$$R_D^{(f)} = max(H2^{(f)}) \tag{6}$$

extracting out maximum values across each filter to obtain the network representations R_C and R_D, respectively:

$$R_C = [R_C^{(1)} \oplus R_C^{(2)} \oplus ..... \oplus R_C^{(f)}]^T \tag{7}$$

$$R_D = [R_D^{(1)} \oplus R_D^{(2)} \oplus ..... \oplus R_D^{(f)}]^T \tag{8}$$

R_C is then merged with S_C and R_D with S_D to get the final representations $F1 = R_C \oplus S_C$ and $F2 = F2 = R_D \oplus S_D$, respectively, which are then passed through two separate multi-perceptron dense (fully connected) layers with regularization (to avoid overfitting). This gives us the two networks DN_C and DN_D:

$$DN_C^{(i)} = \phi(\sum_{j=0}^{(f+4)-1} \varpi_{i,j} \cdot F1_j + \beta) \tag{9}$$

$$DN_D^{(i)} = \phi(\sum_{l=0}^{(f+4)-1} \varpi_{i,l} \cdot F2_l + \beta) \tag{10}$$

where f denotes the number of filters in the convolutional layer; 4 is the length of S_C and S_D; $\varpi_{i,j}$ and $\varpi_{i,l}$ are the weights of the connection between ith neuron of the hidden layer DN_C and DN_D, jth and lth are the outputs of $F1$ and $F2$, respectively; β represents bias; and ϕ is a nonlinear activation function. These layers are finally combined for the softmax binary classification:

$$I^n = \sum_{i=0}^{m-1} \varpi_{k,i} \cdot (DN_C^{(i)} \oplus DN_D^{(i)}) + \beta \tag{11}$$

$$\hat{\gamma_n} = \frac{\exp^{I^n}}{\sum_{j=0}^{1} \exp^{I^j}} \tag{12}$$

where $n = 0, 1$; $\varpi_{k,i}$ is the weight of the connection between kth output neuron and the concatenated output of $DN_C^{(i)}$ and $DN_D^{(i)}$; and $\hat{\gamma_n}$ is the prediction of the stance of a given claim towards a given document. As output class label we consider the one with the highest probability score. The entire network is optimised using Log likelihood cost function L:

$$L = \sum_{n=0}^{1} \gamma_n \ln \hat{\gamma_n} \tag{13}$$

where γ_n is the true target.

3.3 Stage 3: Agree/Disagree Classification

Since the number of documents taking a stance is usually small, especially the number of *disagree* documents, training a deep neural model efficiently is difficult. We hypothesize that a statistical machine learning algorithm trained with a well-defined set of features can be more effective. We again experimented with several models including SVM with and without class-wise penalty, gradient boosting, decision tree, and random forest, as well as with oversampling methods. We found that an SVM classifier similar to the one of stage 1 (depicted in Fig. 3) obtains the best performance. Specifically, we effectively solve the same optimization function, while the model differs from the model of stage 1 in terms of the set of considered features.

As mentioned before, sentiment-related features are useful in the problems related to opinion/stance classification. Thus, similar to stage 2, we again use the sentiment features S_C and S_D (for the claim and document, respectively) generated using NLTK. In addition, we exploit linguistic features generated using the LIWC tool (Linguistic Inquiry and Word Count).[4] LIWC returns more than 90 features related to various linguistic properties of the input text. We selected the following 16 features that seem to be useful for understanding the agreement/disagreement stance: *analytical thinking*, *clout* (expressing confidence in perspective), *authentic*, *emotional tone*, *conjugation*, *negation*, *comparison words*, *affective processes*, *positive emotions*, *negative emotions*, *anxiety*, *anger*, *sadness*, *differentiation* (distinguishing between entities), *affiliation* (references to others), and *achieve* (reference to success, failure). Moreover, we consider the *refuting words* feature set used in the baseline model provided by the FNC-I organizers. This feature set is generated by matching words from a predefined set of *refuting words* with the document's words. The result is a feature vector of the length same as the number of refuting words. The vector contains "1" or "0" in each position i, depending on whether the corresponding refuting word exists in the document or not.

4 Evaluation

4.1 Evaluation Setup

4.1.1 Dataset

We use the benchmark dataset provided by the Fake News Challenge - Stage 1 (FNC-I)[5] [33], which focuses on the same *4-class* stance classification task (of documents towards claims). While there are datasets for related stance detection tasks available, these either address binary or three-class classification problems, or focus on detecting the stance of *user opinions* regarding topics. Thus, they are not suitable to assess the 4-class stance detection problem (of Web documents) addressed in our work.

The FNC-I dataset was derived from the Emergent dataset [15] and consists of 2,587 archived documents related to 300 claims. Each document has a summarised *headline* which reflects the stance of its text (this means that each claim can be represented through different headlines of different stances which make the problem harder). The FNC-I dataset contains 49,972 training and 25,413 test instances, related to 200 and 100 different claims, respectively. Each instance has three attributes: a *headline* (which in our case has the role of a *claim*), ii) a *body text* (document), and iii) a *stance label*, having one of the following values: *unrelated*, *discuss* (*neutral*), *agree*, and *disagree*. The class distribution (Table 1) shows the strong class imbalance: there is a very large number of *unrelated* documents (more than 70% in both training and testing datasets), a large number of *neutral* documents (about 18%), and a very small number of agree ($< 8\%$) and disagree ($< 3\%$) documents.

For the first stage of our pipeline (relevance classification), we merge the classes *discuss*, *agree* and *disagree*, to one *related* class, facilitating a binary *unrelated-related*

[4] `http://liwc.wpengine.com/`

[5] `http://www.fakenewschallenge.org/`

Table 1 Data distribution of the FNC-I dataset.

	All	Unrelated	Neutral	Agree	Disagree
Train	49,972	36,545	8,909	3,678	840
Test	25,413	18,349	4,464	1,903	697

Table 2 Class distribution for *relevance* classification.

	Instances	Unrelated	Related
Train	49,972	36,545	13,427
Test	25,413	18,349	7,064

Table 3 Class distribution for *discuss/stance* classification.

	Instances	Neutral	Stance
Train	13,427	8,909	4,518
Test	7,064	4,464	2,600

Table 4 Class distribution for *agree/disagree* classification.

	Instances	Agree	Disagree
Train	4,518	3,678	840
Test	2,600	1,903	697

classification task. Table 2 documents that *unrelated* documents are more than twice the *related* documents. For the second stage (neutral/stance classification), we consider only *related* documents and merge the classes *agree* and *disagree* to one *stance* class and consider the rest as *neutral*. Table 3 shows the corresponding class distribution. We notice that the *neutral* documents are about twice the *stance* documents. For the final stage (agree/disagree classification), we only consider the instances of the *agree* and *disagree* classes. Table 4 documents that the classes are imbalanced with the disagree class amounting to only 18.6% (26.8%) of the instances in the training (test) dataset.

4.1.2 Evaluation Metrics

The FNC-I task was evaluated based on a weighted (two-level) scoring system which awards 0.25 points if a document is correctly classified as *related* or *unrelated*, and an additional 0.75 points if it is correctly classified as *neutral*, *agree* or *disagree*. However, as argued in [18], this metric fails to take into account the highly imbalanced class distribution of the classes *neutral*, *agree*, *disagree*. For example, a classifier that always predicts *neutral* after a correct *related* prediction achieves a score of 0.833, which is higher than the top-ranked system in FNC-I. Moreover, an effective stance classification model should perform well for the important classes, *agree* and *disagree*, since such documents provide crucial evidence when aiming to detect false claims or validate true claims. On the other hand, the *unrelated* and *neutral* classes are not important in this context. For instance, a document that discusses about a claim without taking a stance is not actually

useful in fact-checking since it does not provide actual evidence about the veracity of the claim.

Based on the above observations, apart from the FNC-I evaluation measure, we also consider the following metrics for the overall assessment and comparison: i) the class-wise F1 score (the harmonic mean of precision and recall for each class), ii) the macro-averaged F1 score across all the four classes ($F1^m$), and iii) the macro-averaged F1 score across the important classes *agree* and *disagree* ($F1^m_{Agr/Dis}$).

4.1.3 Baselines

We consider the below nine baseline methods:

- *Majority vote*: The class with the maximum number of instances is always selected (*unrelated* in our case).
- *FNC baseline*[6]: A gradient boosting classifier using a set of hand-crafted features relevant for the task. The features include word/n-gram overlap features and indicator features for polarity and refutation.
- *SOLAT in the SWEN*[7] [4]: The top-ranked system of FNC-I. This model is based on a weighted average between gradient-boosted decision trees and a deep convolutional neural network. The considered features include: word2vec embeddings, number of overlapping words, similarities between the word count, 2-grams and 3-grams, and similarities after transforming the counts with TF-IDF weighting and SVD.
- *Athene (UKP Lab)*[8] [17]: The second-ranked system of FNC-I, based on a multilayer perceptron classifier (MLP) with six hidden and a softmax layer. It incorporates the following hand-crafted features: unigrams, cosine similarity of word embeddings of nouns and verbs between claim and document, topic models based on non-negative matrix factorization, latent Dirichlet allocation, and latent semantic indexing, in addition to the features provided in the FNC-I baseline.
- *UCL Machine Reading (UCLMR)*[9] [35]: The third-ranked system of FNC-I based on simple MLP network with a single hidden layer. As features it uses the TF vectors of unigrams of the 5,000 most frequent words, and the cosine similarity of the TF-IDF vectors of the claim and document.
- *ComboNSE* [6]: A deep MLP model which combines neural, statistical and external features. Specifically, the model uses neural embeddings from a deep recurrent model, statistical features from a weighted n-gram bag-of-words model, and hand-crafted external features (including TF, ngrams, TF-IDF, and sentiment features).
- *StackLSTM* [18]: A model which combines hand-crafted features (selected through extensive feature analysis) with a stacked LSTM network, using 50-dimensional GloVe word embeddings [32], in order to generate sequences of word vectors of a claim-document pair.
- *LearnedMMD* [47]: A two-layer hierarchical neural network that controls the error propagation between the two layers using a Maximum Mean Discrep-

[6] `https://github.com/FakeNewsChallenge/fnc-1-baseline`
[7] `https://github.com/Cisco-Talos/fnc-1/`
[8] `https://github.com/hanselowski/athene_system`
[9] `https://github.com/uclmr/fakenewschallenge`

ancy regularizer. The first layer distinguishes between the *related* and *unrelated* classes, and the second detects the actual stance. We report the results obtained using the provided code.[10]
– *3-Stage Trad* [27]: A three-stage classification approach similar to our pipeline method that makes use of two traditional classifiers (L1-Regularized Logistic Regression and Random Forest) and a set of 18 features.

We compare the performance of the above baselines with our pipeline method *L3S* (Learning in 3 Steps). Whereas [44] may be considered as an additional baseline, we were not able to compare performance, since the data used in this method is not provided by the authors (such as the used vocabulary for contradiction indicators), and this method disregards the *neutral (discuss)* class used in the FNC-I dataset and our work.

4.1.4 Implementation

We use the Keras deep learning framework [10] with TensorFlow [1] to implement our model in stage 2. For implementation of the models in stage 1 and stage 3 we use the Scikit-learn library [31].

4.2 Evaluation results

4.2.1 Overall classification performance

Table 5 shows the performance of all the approaches. First, we notice that, considering the problematic FNC-I evaluation measure, *CombNSE* achieves the highest performance (0.83), followed by *SOLAT*, *Athene*, *UCLMR*, *StackLSTM*, *3-Stage Trad.* (0.82) and our pipeline approach (0.81). However, we see that the ranking of the top performing systems is very different if we consider the more robust macro-averaged F1 measure ($F1^m$). Specifically, *L3S* now achieves the highest score (0.62), outperforming the best baseline system (*stackLSTM*) by one percentage point, while *CombNSE* is now in the fourth position ($F1^m = 0.59$).

This overall performance gain of *L3S* is, in particular, due to the robustness of the classifier in predicting the *disagree* class, which is particularly difficult to classify due to the low number of training instances. Specifically, our method improves the F1 score of this class by 28% (from 0.18 to 0.23) compared to the best performing baseline for the same class (*stackLSTM*). All other comparing baselines achieve less than 0.15 F1 score for this class. With respect to the *agree* class, *SOLAT* is the top performing system, slightly outperforming our pipeline method by one percentage point of F1 score. However, *SOLAT* performs poorly on the *disagree* class achieving only 3% F1 score. Considering now both *agree* and *disagree*, our pipeline method achieves the highest macro-averaged F1 score ($F1^m_{\text{Agr/Dis}}$), improving the state-of-the-art performance by around 12%.

With respect to the other two classes (*unrelated* and *neutral*), we note that *unrelated* achieves a very high F1 score for all the methods. This is expected given the nature of the classification problem and that the majority of instances belongs to this class. In the *neutral* class, the top performing system (*Athene*) achieves 0.78

[10] `https://github.com/QiangAIResearcher/hier_stance_detection`

Table 5 Document stance classification performance.

System	FNC	$F1^m$	$F1_{\text{Unrel.}}$	$F1_{\text{Neutral}}$	$F1_{\text{Agree}}$	$F1_{\text{Disagr.}}$	$F1^m_{\text{Agr/Dis}}$
Majority vote	0.39	0.21	0.84	0.00	0.00	0.00	0.00
FNC baseline	0.75	0.45	0.96	0.69	0.15	0.02	0.09
SOLAT [4]	0.82	0.58	**0.99**	0.76	**0.54**	0.03	0.29
Athene [17]	0.82	0.60	**0.99**	**0.78**	0.49	0.15	0.32
UCLMR [35]	0.82	0.58	**0.99**	0.75	0.48	0.11	0.30
CombNSE [6]	**0.83**	0.59	0.98	0.77	0.49	0.11	0.30
StackLSTM [18]	0.82	0.61	**0.99**	0.76	0.50	0.18	0.34
LearnedMMD [47]	0.79	0.57	0.97	0.73	0.50	0.09	0.29
3-Stage Trad [27]	0.82	0.59	0.98	0.76	0.52	0.10	0.31
L3S	0.81	**0.62**	0.97	0.75	0.53	**0.23**	**0.38**

Table 6 Class-wise performance of the different pipeline stages and the entire pipeline system.

Stage	Class	P	R	F1
Stage 1	Unrelated	0.97	0.96	0.97
	Related	0.91	0.93	0.92
	Macro-averaged:	0.94	0.95	0.95
Stage 2	Neutral	0.82	0.80	0.81
	Stance	0.67	0.71	0.69
	Macro-averaged:	0.75	0.76	0.75
Stage 3	Agree	0.79	0.75	0.77
	Disagree	0.40	0.44	0.42
	Macro-averaged:	0.60	0.60	0.60
Pipeline	Unrelated	0.97	0.96	0.97
	Neutral	0.74	0.76	0.75
	Agree	0.52	0.53	0.53
	Disagree	0.22	0.23	0.23
	Macro-averaged:	0.61	0.62	0.62

F1 score while our approach gives 0.75. Note however that, as we have already argued, similar to the *unrelated* class the *neutral* class is not usually useful in a fact checking context.

4.2.2 Detailed per-stage performance of L3S

We now study the performance of each stage separately as well as the detailed performance of our pipeline system in terms of per-class and macro-averaged precision (P), recall (R) and F1 score (Table 6).

With respect to stage 1 of our pipeline (relevance classification), precision and recall of the *related* class (the important class is this stage) is 0.91 and 0.93, respectively. Although the data is very imbalanced as shown in Table 2 (*unrelated* corresponds to around 28% of all test instances), performance is comparably high.

Regarding stage 2 (neutral/stance classification), precision and recall of the important *stance* class is 0.67 and 0.71, respectively, while that of the *neutral* class is 0.82 and 0.80, respectively. In general, this task is harder than relevance classification (stage 1). Here, again the data is imbalanced, with the *stance* instances

Table 7 Confusion matrix of our pipeline system.

	Agree	Disagree	Neutral	Unrelated
Agree	1,006	278	495	124
Disagree	237	160	171	129
Neutral	555	252	3,381	276
Unrelated	127	31	523	17,668

corresponding to around 37% of the test instances (see Table 3). We note that there is room for improvement for the important *stance* class.

The stage 3 of our pipeline deals with the harder problem of agree/disagree classification. We notice that our classifier performs well on the *agree* class (P = 0.79, R = 0.75), but poorly on the *disagree* class (P = 0.40, R = 0.44). We observe that, even a dedicated classifier which only considers *stance* documents struggles detecting many instances of the *disagree* class. Nevertheless, our method outperforms the existing methods by more than 28% of F1 score. There are two main reasons affecting the performance of the *disagree* class: i) the classifiers of stage 2 and 3 may filter out many instances belonging to this class, ii) the amount of training instances is limited for this class (only 840), while the data distribution is very imbalanced with the *disagree* class corresponding to 18.6% of the test instances (as shown in Table 4). Regarding the latter, applying oversampling methods [9] for coping with the limited amount of training instances for the minority *disagree* class did not improve the performance in our experiments.

Observing precision and recall of each class for the whole pipeline approach and comparing these values with the values of the same classes in each different stage, illustrates the effects of the filtering process applied in each stage on the performance of the next stage(s). For instance, we observe that recall of the *disagree* class is 0.44 in stage 3 but only 0.23 overall. The same problem exists for the *agree* class (from 0.75 to 0.53). This suggests that many *stance* documents are misclassified in the previous stages, making the task of stage 3 harder. We try to better understand this problem through an error analysis in the following subsection.

4.2.3 Error analysis

Table 7 shows the confusion matrix of our pipeline system. Overall, we note that the *neutral* and *agree* classes seem to be frequently confused, what seems intuitive given the very similar nature of these classes, i.e. a document which discusses a claim without explicitly taking a stance is likely to agree with it. The results for the *disagree* class illustrate that stage 1 misclassifies (as *unrelated*) 18.5% of all *disagree* instances (129 instances, in total), while this percentage is less than 7% for the *agree* and *neutral* classes. In stage 2, we see that 171 *disagree* instances (25.5%) are misclassified as *neutral*, while this percentage is similar for the *agree* class (26.2%). Finally, in the last stage, 34% of the *disagree* instances are misclassified as *agree*, which demonstrates the difficulty of this task. As we explained above, the highly unbalanced data distribution and the limited amount of training data are likely to contribute significantly to this picture.

Table 8 shows the confusion matrix of each stage separately, i.e., without the effect of the filtering process. We note the increasing difficulty of each stage. Stage

Table 8 Confusion matrix of different stages.

		Unrelated	Related
Stage1	Unrelated	17,668	681
	Related	529	6,535
		Neutral	Stance
Stage2	Neutral	3,575	889
	Stance	760	1,840
		Agree	Disagree
Stage3	Agree	1,436	467
	Disagree	387	310

1 misclassifies a small number of *related* instances as *unrelated* (less than 8%). Stage 2 misclassifies 29% of the *stance* instances as *neutral*. Finally, stage 3 misclassifies the majority of *disagree* instances (55%) as *agree*, and around 25% of the *agree* instances as *disagree*. It is evident from these results that there is much room for improvement for the last stage of our pipeline.

To better understand the misclassification problem, we further analyze several misclassified *disagree* instances. We observe that the majority of cases concern a small number of distinct claims. In stage 1, for instance, 56/129 (43.4%) misclassifications are associated with one claim (the claim *"Florida woman underwent surgery to add a third breast"*). The instances of this claim were misclassified as *unrelated* because of vocabulary mismatch: different words are used to express *"breast"* in the claim and the documents (*"boob"*, *"breast"*), and the distance of these words in the embedded vector is unexpectedly high. Similarly, in stages 2 and 3, there are 40/171 (23.4%) and 49/237 (20.7%) misclassification cases, respectively, associated with only one claim. In stage 2 and 3, the reason is mainly due to the lack of semantic understanding of the text used to express the disagreement to the claim. For example, a misclassified document uses the phrase *"shot down a report claiming..."* in the 2nd paragraph, while the remaining part of the document does not discuss about the claim itself. Another example of vocabulary mismatch is a misclassified document uses the text *"all of it is bullsh*t"*, and *"it is a nice mixture of folklore and truth"*. In these cases, the model fails to understand that the words *bullsh*t* and *folklore* negate the claim.

5 Related Work

Stance detection is a classification problem in natural language processing where the stance of a *(piece of) text* towards a particular *target* is explored. Stance detection has been applied in different contexts, including *social media* (stance of a tweet towards an entity or topic) [30,13,3,24,40,14,46], *online debates* (stance of a user post or argument/claim towards a controversial topic or statement) [43, 39,5,16], and *news media* (stance of an article towards a claim) [33,18,6,44,47]. Our work falls under the context of *news media* where the ultimate objective is the detection of fake news. Below we discuss related works on this area and the difference of our approach.

The 4-class stance classification problem for news media was introduced in the context of the Fake News Challenge (FNC) [33], as *"a helpful first step towards identifying fake news"*.[11] The organizers made available a ground truth dataset as well as a simple baseline method that uses a set of hand-coded features and a gradient boosting classifier. 50 teams participated in FNC using a wide array of techniques and sets of hand-crafted features. We consider the top-3 performing systems as baselines in our experiments (more in Section 4.1).

For the same problem, [29] demonstrated the challenges to apply conditional encoding and attention recurrent networks, methods known to work well in other stance detection problems. [6] proposed a deep MLP model with combined neural, statistical, and external features, achieving a state of the art performance. [18] proposed the use of macro-averaged F1 score for evaluating performance in this task because this metric is less affected by highly imbalanced datasets. Moreover, the authors proposed a novel approach which uses a stacked LSTM network and 50-dimensional GloVe word embeddings [32], outperforming all previous methods on macro-averaged F1 score.

All these methods treat the problem as a single 4-class classification task. [44] proposed a different, two-stage approach where *unrelated* documents are first filtered out through relevance classification, and then the *related* documents are classified as *contradict* (*disagree*) or *support* (*agree*), using a gradient boosted decision tree model trained on n-gram features extracted using a specially-designed *contradiction vocabulary*. However, this model ignores the *discuss* class which is very prominent in fake news. Moreover, the authors do not provide the evaluation datasets and the used contradiction vocabulary. A more recent work [47] also applies a two-stage approach where a first stage distinguishes *related* from *unrelated* documents and a second detects the actual stance (*agree*, *disagree*, *neutral*). A hierarchical neural network that controls the error propagation between the two stages has been proposed. We have not been able to replicate the results as reported in [47] after using the provided code. Finally, [27] proposes a three-stage approach similar to our proposed architecture that only makes use of traditional machine learning models (L1-Regularized Logistic Regression and Random Forest) and a large number of features. However, this method also fails to achieve a satisfactory performance on the important disagree class (see Section 4.2).

With respect to evaluation datasets, a range of datasets has been made available for related stance classification problems, for instance, for detecting the stance of ideological debates (for/against the debate topic) [19], context-dependent arguments/claims (pro/con a controversial statement) [5], or tweets (favor/against a controversial topic) [30]. However, the ground truth dataset provided by the FNC-I is, to the best of our knowledge, the only one focusing on the *4-class* stance classification task of Web documents addressed in this paper. Also, the aforementioned datasets focus on detecting the stance of *user opinions* regarding topics, as opposed to the stance of Web documents (like news articles) towards a given true or false claim.

Finally, there is a number of complementary research works that are not directly concerned with stance detection but which tackle problems relevant to fact checking and fake news detection, such us trustworthiness assessment [38], spam-

[11] `http://www.fakenewschallenge.org/`

mer detection [7,12], detection of check-worthy claims [20], or verified claim retrieval [37].

6 Conclusion

We have proposed a modular pipeline of cascading classifiers for the problem of document stance classification towards claims. This enables the use of different classifiers and features in each pipeline step, crucial to further optimize performance on a per step and class basis. Experimental results on the benchmark dataset demonstrated the state-of-the-art performance of our pipeline model and its ability to improve the performance of the important—for fake news detection—*disagree* class by 28% (F1 score) without significantly affecting the performance of the *agree* class. The results also showed that there is still room for further improvements, mainly due to the lack of semantic understanding of the language used to express agreement or disagreement. As part of future work, we are planning to focus on this problem, aiming to reduce the number of misclassified *agree* and *disagree* documents of each stage.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), pp. 265–283 (2016)
2. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. Journal of Economic Perspectives **31**(2), 211–36 (2017)
3. Augenstein, I., Rocktäschel, T., Vlachos, A., Bontcheva, K.: Stance detection with bidirectional conditional encoding. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 876–885. ACL (2016). DOI 10.18653/v1/D16-1084. URL `http://aclweb.org/anthology/D16-1084`
4. Baird, S., Sibley, D., Pan, Y.: Talos targets disinformation with fake news challenge victory. `https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html` (2017)
5. Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., Slonim, N.: Stance classification of context-dependent claims. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 251–261 (2017)
6. Bhatt, G., Sharma, A., Sharma, S., Nagpal, A., Raman, B., Mittal, A.: Combining neural, statistical and external features for fake news stance identification. In: Companion Proceedings of the The Web Conference 2018, WWW '18, pp. 1353–1357. International World Wide Web Conferences Steering Committee (2018). DOI 10.1145/3184558.3191577. URL `https://doi.org/10.1145/3184558.3191577`
7. Bindu, P., Mishra, R., Thilagam, P.S.: Discovering spammer communities in twitter. Journal of Intelligent Information Systems **51**(3), 503–527 (2018)
8. Chang, J., Lefferman, J., Pedersen, C., Martz, G.: When fake news stories make real news headlines. Nightline. ABC News. `https://abcnews.go.com/Technology/fake-news-stories-make-real-news-headlines/story?id=43845383` (2016)
9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
10. Chollet, F., et al.: Keras. `https://keras.io` (2015)
11. Conforti, C., Pilehvar, M.T., Collier, N.: Towards automatic fake news detection: Cross-level stance detection in news articles. In: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pp. 40–49 (2018)

12. Dewang, R.K., Singh, A.K.: State-of-art approaches for review spammer detection: a survey. Journal of Intelligent Information Systems **50**(2), 231–264 (2018)
13. Du, J., Xu, R., He, Y., Gui, L.: Stance classification with target-specific neural attention. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, pp. 3988–3994 (2017). DOI 10.24963/ijcai.2017/557. URL https://doi.org/10.24963/ijcai.2017/557
14. Ebrahimi, J., Dou, D., Lowd, D.: Weakly supervised tweet stance classification by relational bootstrapping. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1012–1017 (2016)
15. Ferreira, W., Vlachos, A.: Emergent: a novel data-set for stance classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies, pp. 1163–1168 (2016)
16. Guggilla, C., Miller, T., Gurevych, I.: Cnn-and lstm-based claim classification in online user comments. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 2740–2751 (2016)
17. Hanselowski, A., Avinesh, P., Schiller, B., Caspelherr, F.: Description of the system developed by team athene in the fnc-1. Tech. rep., Technical report (2017)
18. Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C.M., Gurevych, I.: A retrospective analysis of the fake news challenge stance-detection task. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1859–1874. ACL (2018). URL http://aclweb.org/anthology/C18-1158
19. Hasan, K.S., Ng, V.: Stance classification of ideological debates: Data, models, features, and constraints. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 1348–1356 (2013)
20. Hassan, N., Arslan, F., Li, C., Tremayne, M.: Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1803–1812 (2017)
21. Hassan, N., Li, C., Tremayne, M.: Detecting check-worthy factual claims in presidential debates. In: Proceedings of the 24th acm international on conference on information and knowledge management, pp. 1835–1838. ACM (2015)
22. Hutto, C.J., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: ICWSM (2014)
23. Küçük, D., Can, F.: Stance Detection: A Survey. ACM Computing Surveys **53**(1), 12:1–12:37 (2020). DOI 10.1145/3369026. URL https://doi.org/10.1145/3369026
24. Lai, M., Hernández Farías, D.I., Patti, V., Rosso, P.: Friends and enemies of Clinton and Trump: Using context for detecting stance in political tweets. In: G. Sidorov, O. Herrera-Alcántara (eds.) Advances in Computational Intelligence, pp. 155–168. Springer International Publishing, Cham (2017)
25. Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., et al.: The science of fake news. Science **359**(6380), 1094–1096 (2018)
26. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision, pp. 2999–3007 (2017)
27. Masood, R., Aker, A.: The fake news challenge: Stance detection using traditional machine learning approaches. In: KMIS, pp. 126–133 (2018)
28. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, pp. 3111–3119. Curran Associates Inc., USA (2013)
29. Miller, K., Oswalt, A.: Fake news headline classification using neural networks with attention. Tech. rep., tech. rep., California State University, year (2017)
30. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: Semeval-2016 task 6: Detecting stance in tweets. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 31–41 (2016)
31. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
32. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543 (2014)

33. Pomerleau, D., Rao, D.: Fake news challenge stage 1 (FNC-I): Stance detection. http://www.fakenewschallenge.org/ (2017)
34. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: Credibility assessment of textual claims on the web. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 2173–2178. ACM (2016)
35. Riedel, B., Augenstein, I., Spithourakis, G.P., Riedel, S.: A simple but tough-to-beat baseline for the fake news challenge stance detection task. arXiv preprint arXiv:1707.03264 (2017)
36. Ruder, S., Glover, J., Mehrabani, A., Ghaffari, P.: 360° stance detection. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 31–35. Association for Computational Linguistics (2018). DOI 10.18653/v1/N18-5007. URL http://aclweb.org/anthology/N18-5007
37. Shaar, S., Babulkov, N., Da San Martino, G., Nakov, P.: That is a known lie: Detecting previously fact-checked claims. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3607–3618 (2020)
38. Sherchan, W., Nepal, S., Paris, C.: A survey of trust in social networks. ACM Computing Surveys (CSUR) **45**(4), 1–33 (2013)
39. Sridhar, D., Foulds, J., Huang, B., Getoor, L., Walker, M.: Joint models of disagreement and stance in online debate. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp. 116–125. ACL (2015). DOI 10.3115/v1/P15-1012. URL http://aclweb.org/anthology/P15-1012
40. Sun, Q., Wang, Z., Zhu, Q., Zhou, G.: Stance detection with hierarchical attention network. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 2399–2409. ACL (2018). URL http://aclweb.org/anthology/C18-1203
41. Veropoulos, K., Campbell, C., Cristianini, N.: Controlling the sensitivity of support vector machines. In: Proceedings of the International Joint Conference on AI, pp. 55–60 (1999)
42. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
43. Walker, M.A., Anand, P., Abbott, R., Grant, R.: Stance classification using dialogic properties of persuasion. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT '12, pp. 592–596. ACL (2012)
44. Wang, X., Yu, C., Baumgartner, S., Korn, F.: Relevant document discovery for fact-checking articles. In: Proceedings of the 2018 World Wide Web Conference, pp. 525–533. International World Wide Web Conferences Steering Committee, Lyon, France (2018)
45. Wu, Y., Agarwal, P.K., Li, C., Yang, J., Yu, C.: Toward computational fact-checking. Proceedings of the VLDB Endowment **7**(7), 589–600 (2014)
46. Xu, C., Paris, C., Nepal, S., Sparks, R.: Cross-target stance classification with self-attention networks. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 778–783 (2018)
47. Zhang, Q., Liang, S., Lipani, A., Ren, Z., Yilmaz, E.: From stances' imbalance to their hierarchical representation and detection. In: Companion Proceedings of the The Web Conference (2019)