Gradual Argumentation Evaluation for Stance Aggregation in Automated Fake News Detection

Neema Kotonya and **Francesca Toni**
Department of Computing
Imperial College London, United Kingdom
{n.kotonya18,f.toni}@imperial.ac.uk

Abstract

Stance detection plays a pivot role in fake news detection. The task involves determining the point of view or stance – for or against – a text takes towards a claim. One very important stage in employing stance detection for fake news detection is the aggregation of multiple stance labels from different text sources in order to compute a prediction for the veracity of a claim. Typically, aggregation is treated as a credibility-weighted average of stance predictions. In this work, we take the novel approach of applying, for aggregation, a gradual argumentation semantics to bipolar argumentation frameworks mined using stance detection. Our empirical evaluation shows that our method results in more accurate veracity predictions.

1 Introduction

The problem of fake news has existed from time immemorial. But in recent times, both the rise of social media as the go-to platform for receiving news updates and a series of significant politic elections events, the results of which are speculated to have been influenced by misinformation, has culminated in the phrase being pushed to the forefront of our consciousness. It is widely acknowledged (e.g., see (Lazer et al., 2018)) that fake news is an important problem, and that attention should be directed to tackle it.

Fake news is a particularly challenging problem, one that consists of a number of subproblems, and one for which many approaches have been proposed (e.g., see (Zhou et al., 2019)). Generally fake news detection amounts to collating evidence and counter-evidence from various sources in order to make an assessment regarding the veracity of a given claim, e.g., as in the Fact Extraction and Verification (FEVER) shared task (Thorne et al., 2018).

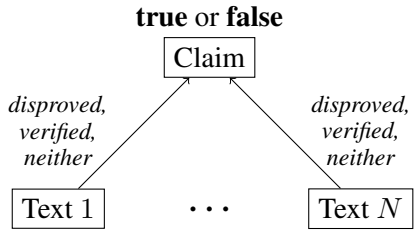


Figure 1: In veracity assessment a true/false label can be acquired by aggregating various texts that *verified* and *disproved* the target claim.

Veracity assessment is typically formulated as a 3-class problem where we aim to arrive at a value for the factuality of a claim, which is based on the stances of Texts $1, \ldots, N$ (see Figure 1). These texts could be headlines, articles, and even other claims. One of the tasks underpinning the prediction of factuality is stance detection. It involves examining agreement expressed by a text in relation to a claim. The text could be a headline (Ferreira and Vlachos, 2016), a topic (Mohammad et al., 2016) or a lengthier text fragment (Pomerleau and Rao, 2017). Stance detection can be thought of as a two-part task: we first aim to determine if the text and claim are sufficiently close with respect to their subject matter, and then, once relatedness of the text and claim is established, we want to know whether the text takes a favourable or unfavourable view of the claim.

The intuition behind the use of stance detection for fake news analysis is that the trustworthiness of a claim is strongly tied to the level of agreement expressed either for or against it in other texts, particularly the agreement or disagreement expressed by sources with high credibility. For that reason, we should be able to aggregate these disjoint stance valuations in order to arrive at a prediction for the veracity of the claim, as described by Conforti et al. (2018).

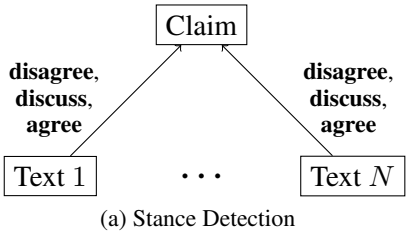
Proceedings of the 6th Workshop on Argument Mining, pages 156–166
Florence, Italy, August 1, 2019.

In this paper we draw inspiration from uses of relation-based argument mining (Carstens and Toni, 2015) to generate and evaluate bipolar argumentation frameworks (BAFs) (Cayrol and Lagasquie-Schiex, 2005) in order to perform classification tasks (e.g., in (Cocarascu and Toni, 2018), for deception detection). In the same spirit, we propose and use a stance detection classifier to generate BAFs and evaluate arguments therein with the existing DF-QuAD gradual semantics (Rago et al., 2016) in order to assess veracity of news against evidence. We show empirically, using a stance detection classifier built from the Fake News Challenge dataset (Pomerleau and Rao, 2017) and tested on the RumourEval dataset (Derczynski et al., 2017), that DF-QuAD performs competitively in comparison with a standard stance aggregation method using a credibility-weighted average of stance predictions. The aggregation method resulting from deploying DF-QuAD, unlike the standard aggregation method, considers also the dialectical relationships between different evidence and counter-evidence texts in order to gauge the veracity of target claims.

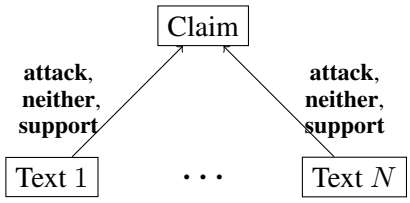
2 Related Work

Stance detection can be framed as a four-way classification problem, as in the Fake News Challenge (Pomerleau and Rao, 2017), where it is aimed at identifying, in pairs consisting of headlines and article bodies, whether the texts are UNRELATED, or if the article body AGREES, DISAGREES, or DISCUSSES the headline. The last label signifies that the two texts are related but no stance (for or against) exists from the body to the headline. The RumourEval rumour verification task in SemEval 2017 (Derczynski et al., 2017) similarly includes a stance detection sub-task and uses data in the format of pairs but labels stances as DENY, SUPPORT, COMMENT and QUERY. In this paper, we see stance detection as a three-way classification problem, as summarized in Figure 2(a), assuming that relatedness has already been ascertained. This is in line with other work, notably in the EMERGENT project[1], using three labels FOR, AGAINST and OBSERVING (Ferreira and Vlachos, 2016).

Given the almost parallel stance labels, when restricted to three, between Fake News Challenge and RumourEval, we choose to develop classifiers for stance detection using the former and verify them on the latter, for veracity prediction.



(a) Stance Detection



(b) Relation-based Argument Mining

Figure 2: Labels for relation-based argument mining, stance detection and veracity assessment. The labels in bold are those learnt from the task.

A number of techniques have been employed for the purpose of building stance detection systems (Hanselowski et al., 2018), including Long-Short Term Memory networks (LSTMs) (Hanselowski et al., 2018; Shang et al., 2018), term frequency-inverse document frequency (TF-IDF) and bag of word (BOW) features with Multi-Layer Perceptrons (Riedel et al., 2017), end-to-end memory networks enhanced with CNNs and LSTMs (Mohtarami et al., 2018), and non-neural network and neural network classifiers using cue words, Google News word2vec embeddings, and features taken from the Fake News Challenge dataset (Ghanem et al., 2018). We experiment with gradient-boosting, Gated Recurrent Units (GRUs), LSTMs and bidirectional LSTMs (BiLSTMs).

In terms of label aggregation for veracity assessment, Popat et al. (2018) derive credibility assessments for text-based claims aggregating a number of web-sourced articles. Source embeddings for both claims and articles are used to weigh the claims' credibility, and are derived from the names of sources who published the claims e.g., news organisations as well as individuals, typically public figures such as politicians. In this paper we perform aggregation using a gradual semantics for bipolar argumentation (see Section 3), taking into account the stance of responses towards claims and other responses.

[1] http://www.emergent.info/about

Gradual semantics and bipolar argumentation for classification have been used for other tasks, notably in (Cocarascu and Toni, 2018) to contribute features for detecting deceptive reviews. There, bipolar argumentation frameworks were obtained using relation-based argument mining, as understood in (Carstens and Toni, 2015) and summarized in Figure 2(b). In this paper, we perform relation-based argument mining by way of stance detection: when stance detection is modelled as a three-class problem, the labels FOR, AGAINST and OBSERVING bear a strong resemblance to ATTACK, SUPPORT and NEITHER considered in relation-based argument mining (Carstens and Toni, 2015). Thus, we use stance relations as argumentative attack and support relations to evaluate the veracity of claims.

Other forms of argument mining have been studied in conjunction with stance detection. These include argument tagging for insufficiently labelled corpora (Sobhani et al., 2015) and identification of argumentative components in social media conversations (Boltužić and Šnajder, 2014).

3 Background

Our method relies on Bipolar Argumentation Frameworks (Cayrol and Lagasquie-Schiex, 2005) for representing the argumentative relations (disagree and agree) between text pairs, and the Discontinuity-Free Quantitative Argumentation Debates (DF-QuAD) algorithm (Rago et al., 2016) for aggregating the strengths of claims according to these relations. A Bipolar Argumentation Framework (BAF) is the triple $\langle Args, R^-, R^+ \rangle$, in which $Args$ is a set of entities, called arguments, and R^- and R^+ are binary attack and support relations between arguments respectively. The BAF with $Args = \{A_1, A_2, A_3, A_4, A_5\}$, attack relation $R^- = \{(A_1, A_2), (A_2, A_1), (A_2, A_3)\}$ and support relation $R^+ = \{(A_4, A_2), (A_4, A_5)\}$ is shown graphically in Figure 3. Note that the A_i can be instantiated in a number of different ways. For this work, we model claims and counter-claims from the RumourEval dataset as arguments. We identify attack and support relations with the help of stance detection.

Various semantics have been proposed for evaluating the dialectical strength of arguments in BAFs. We use the DF-QuAD algorithm originally defined for QuAD frameworks, which are BAFs

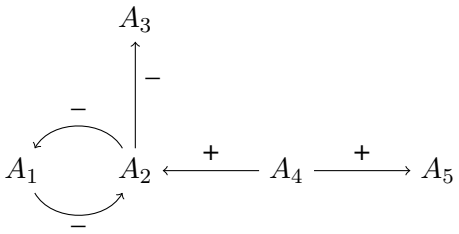


Figure 3: Example BAFs.

$\langle Args, R^-, R^+ \rangle$ forming acyclic graphs with, in addition, each argument $A \in Args$ being attributed a *base score* $\tau(A)$ that denotes its intrinsic strength (prior to considering its attackers $R^-(A) = \{B \in Args | (B, A) \in R^-\}$ and supporters $R^+(A) = \{B \in Args | (B, A) \in R^+\}$).

As required by DF-QuAD, base scores and dialectical strength of arguments are from within $\mathbb{I} = [0, 1]$. In all our experiments, $\tau(A) = 0.5$ for all $A \in Args$. DF-QuAD computes dialectical strength

$$\sigma(A) = \mu(\tau(A), \alpha(\sigma(R^-(A)), \alpha(\sigma(R^+(A)))$$

where $\sigma(R^-(A))$ is the sequence $(\sigma(B_1), \ldots, \sigma(B_n)$ for $R^-(A) = \{B_1, \ldots, B_n\}$, $n \geq 0$ (similarly for $\sigma(R^+(A))$), $\alpha(()) = 0$, $\alpha((v_1)) = v_1$, $\alpha((v_1, v_2)) = f(v_1, v_2) = v_1 + v_2 - v_1 * v_2$ and, for $n > 2$, $\alpha((v_1, \ldots, v_n)) = f(\alpha((v_1, \ldots, v_{n-1}), v_n))$, and, finally, the mediating function $\mu : \mathbb{I} \times \mathbb{I} \rightarrow \mathbb{I}$ is defined as $\mu(v_0, v_a, v_s) = v_0 - v_0 * |v_s - v_a|$ if $v_a \geq v_s$, and $\mu(v_0, v_a, v_s) = v_0 + (1 - v_0) * |v_s - v_a|$otherwise. Intuitively, μ represents the idea that attackers of greater combined strength (given by v_a) than the supporters' combined strength (given by v_s) will weaken an argument (with base score v_0) more severely, i.e., these attackers will bring the argument's strength closer to 0. Similarly, supporters of greater combined strength will bring the argument's strength closer to 1. Conversely, the weaker the attackers or supporters, the smaller the effect on the argument's strength.

By employing DF-QuAD for veracity prediction we make the assumption, for example, that false claims will be weakened by the strength and number of their attackers, and thus have a low dialectical strength as computed using the algorithm, because of their lack of supporting arguments and abundance of attackers. However, we are aware that this might not always be the case, given the presence of silos or echo chambers in

social media. Indeed, in echo chambers fallacious arguments may be backed up by a number of equally misleading arguments, which would result in a high DF-QuAD strength, despite the evidently false claim.

4 Datasets

Two datasets are employed as part of this study: the Fake News Challenge dataset[2], used to train the stance detection classifiers, and the RumourEval dataset[3], which we adapt for the problem of fake news detection to evaluate our argumentation-based stance aggregation methods.

DATASET	CLASS	SIZE (TRAIN+DEV)	
FNC-1	AGREE	3,678	7.36%
	DISAGREE	840	1.68%
	DISCUSS	8,909	17.8%
	UNRELATED	36,545	73.1%
	ALL	49,972	-
RumourEval Task A	COMMENT	2,907	64.3%
	DENY	344	7.61%
	QUERY	358	7.92%
	SUPPORT	910	20.1%
	ALL	4,519	-
RumourEval Task B	FALSE	62	20.9%
	TRUE	137	46.1%
	UNVERIFIABLE	98	33.0%
	ALL	297	-

Table 1: Summary of FNC-1, RumourEval Task A and RumourEval Task B datasets.

4.1 Fake News Challenge

The Fake News Challenge (FNC-1) is a shared task first presented in 2017 for claim verification in the context of news article headlines using machine learning classifiers. Participating groups in the shared task were granted access to training and development datasets consisting of almost 50K examples of headline and article body pairs.

The stance detection task is composed of two sub-problems. First, a classifier must determine if the input texts are related. If relatedness is established, the classifier must then determine whether the article expresses a positive stance (AGREE), a negative stance (DISAGREE), or no stance (DISCUSS) towards the accompanying headline. The following is a truncated example from FNC-1:

Headline: *Spider burrowed through tourist's stomach and up into his chest.*
Article body: *Fear not arachnophobes, the story of Bunbury's "spiderman" might not be all it seemed. Perth scientists have cast doubt over claims that a spider burrowed into a man's body during his first trip to Bali. The story went global on Thursday, generating hundreds of stories online... a specialist dermatologist was called in and later used tweezers to remove what was believed to be a "tropical spider". But it seems we may have all been caught in a web... of misinformation. Arachnologist Dr Volker Framenau said whatever the creature was, it was "almost impossible" for the culprit to have been a spider...*
Label: DISAGREE.

As shown in Table 1, UNRELATED examples account for a large majority (almost three quarters) of the dataset. We discount the UNRELATED label to focus on the three-way classification task of predicting the stance. Thus, we are left with 13,427 examples.

4.2 RumourEval Task A and Task B

Task 8 of SemEval 2017 focused on verifying rumours pertaining to a number of tweets regarding eight contentious topics from current events, captured in the RumourEval dataset, adapted from the PHEME project[4]. The dataset consists of 297 Twitter conversation threads (the English portion of the PHEME journalism use case data). Rumour verification differs from fake news detection in that rumours are not necessarily presented in the form of traditional news media (e.g., newspapers), but the two tasks are related in that they both require the verification of text-based claims.

We were motivated to use the RumourEval dataset because it is annotated for both stance and veracity. Therefore, even though the original SemEval shared task was not formulated with this problem in mind, this dataset is incredibly well-suited to investigating the relation between stance and veracity. Stance (Task A) and veracity (Task B) labels are provided for each of the 297 Twitter threads in the RumourEval dataset (see Table 1). In total this amounts to 4,161 source tweet and

[2] https://github.com/FakeNewsChallenge/fnc-1

[3] http://alt.qcri.org/semeval2017/task8/

[4] https://www.pheme.eu/

reply tweet pairs, once we disregard the QUERY stance detection label. Furthermore, we adapt the remaining stance detection labels, renaming DENY as DISAGREE, SUPPORT as AGREE, and COMMENT as DISCUSS to match the FNC-1 stance labels. As for the veracity labels, we only consider the TRUE and FALSE source tweets. The following text is an excerpt from a conversation thread in the RumourEval dataset regarding the Sydney siege rumour topic:

> **u1/source tweet**: *Up to 20 held hostage in Sydney Lindt Cafe siege ⟨URL⟩ ⟨URL⟩* [SUPPORT]
>
> —**u2/reply 1**: *"@**u1**: Up to 20 held hostage in Sydney Lindt Cafe siege ⟨URL⟩ ⟨URL⟩."* [SUPPORT]
>
> —**u3/reply 2**: *Sick. "@**u1**: Up to 20 held hostage in Sydney Lindt Cafe siege ⟨URL⟩ ⟨URL⟩"* [SUPPORT]
>
> —**u4/reply 3**: **@u1 @u10** *oh god !!!!* [COMMENT]
>
> —**u5/reply 4**: **@u1** *at least they've got good chocolate* [COMMENT]
>
> —**u6/reply 5**: **@u5** *you are an insensitive idiot!* [COMMENT]
>
> —**u7/reply 6**: **@u1** *all reports say 13* [DENY]
>
> —**u8/reply 7**: *"@**u1**: Up to 20 held hostage in Sydney Lindt Cafe siege ⟨URL⟩ ⟨URL⟩" - wonder if they'll get paid overtime* [COMMENT]
>
> —**u9/reply 8**: *"@**u1**: Up to 20 held hostage in Sydney Lindt Cafe siege ⟨URL⟩ ⟨URL⟩" - Oh. My. God. I am SICK!* [COMMENT]
>
> **Task A label**: See conversation thread.
>
> **Task B label**: FALSE

In the above example, the level of indentation is used to distinguish between direct and nested replies. Note that user *u10* does not post a response in the conversation thread, but is *tagged* in the conversation by *u4*. Source tweets also have stance labels relating to whether they support the rumour topic which they concern. Each conversation thread in the RumourEval dataset is accompanied by details pertaining to the conversation structure. This provides information about how the tweets relate to each other, including which are direct replies (e.g., reply 1) and which are nested replies (e.g., reply 5) to the source tweet. We use this structure to construct BAFs.

5 Methodology & Experimental Setup

Our methodology is shown in Figure 4. We train a number of stance detection classifiers on the FNC-1 dataset, the best of which we use to predict the labels for the RumourEval Task A dataset. We then perform stance aggregation on the predicted labels, in order to arrive at a veracity prediction. We compare their veracity assessment performance against the gold standard labels from the RumourEval task B dataset. This allowed us to compare and evaluate the usefulness of the stance detection predictions. The reliability of these labels also enabled us to gauge the effectiveness of stance detection as a tool for veracity assessment.

In the remainder of this section, first we describe the methods we employ for stance classification and then our stance aggregation methods. We developed our own stance detection classifiers using gradient boosting as well as (three forms of) neural networks, of which we selected two (LSTM and BiLSTM) as best performing in stance prediction, to generate BAFs. For stance aggregation, a credibility-weighted average, DF-QuAD with only direct replies, and DF-QuAD with both direct and nested replies, applied to appropriately constructed BAFs using the stance detection classifiers.

5.1 Stance Classification

We implemented four stance detection classifiers. Three of these are recurrent neural networks (RNNs) or bidirectional RNNs (GRU, LSTM, BiLSTM), constructed using the Tensorflow[5] and Keras[6] deep learning libraries. A summary of the hyper-parameters selected for our RNN models is shown in Table 2. We also used a non-neural technique, i.e., gradient boosting. We built the gradient boosting classifier using the Scikit-Learn library module for ensemble classifiers[7].

5.1.1 Preprocessing

All four classifiers were trained using headline-article text pairs extracted from the FNC-1 dataset. The effectiveness of the classifiers was tested on the RumourEval Task A dataset. Note that FNC-1 deals with headlines and article bodies, which are more structured than the tweets which make up the RumourEval dataset, so particular care had

[5] https://tensorflow.org
[6] https://keras.io/
[7] https://scikit-learn.org

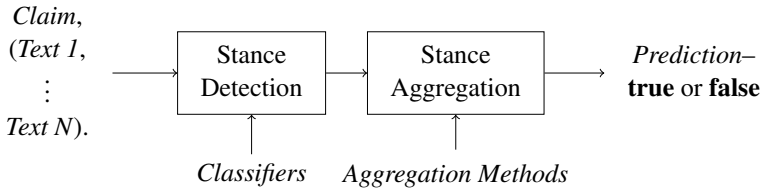


Figure 4: Veracity prediction work flow. As stance classifiers we use LSTMs and BiLSTMs. Methods employed for aggregation are a credibility-weighted average baseline, DF-QuAD (only direct replies), and DF-QuAD (both direct and nested replies).

to be taken in addressing these differences for classifiers trained on the former to perform well when evaluated on the latter.

We used regular expressions to remove links and user handles from tweets. We opted to use 100D pre-trained GloVe embeddings (Pennington et al., 2014) to represent the text inputs. For the deep neural network architectures, we constructed embedding layers. In order to train the non-neural classifier, we computed a mean of each embedding.

We attempted to minimize out-of-vocabulary (OOV) words with lemmatization where possible. Furthermore, we utilized the Stanford Named Entity Recognizer[8] to construct named entity substitutions for locations, organizations, and named people to both minimize OOV words and also prevent over-fitting due to coincidental correlations between named entities and stance labels in the training set, as adopted by Conforti et al. (2018) and Lee et al. (2018). The purpose of employing these techniques was to train more generalized classifiers that would output more accurate predictions when applied to the unseen examples in the RumourEval dataset. This was particularly important given the differences in topics between FNC-1 and RumourEval, but also because FNC-1 contains text pertaining to news articles written in formal English, whereas the RumourEval corpus is composed of short snippets of user-generated text made up of colloquialisms and neologisms which word embeddings is not able to capture semantically.

Furthermore, we made the choice to use stratified cross-validation for training the classifiers. This was because, as can be seen in Table 1, the FNC-1 dataset is highly unbalanced. Although we performed 3-way classification to learn the AGREE, DISAGREE, and DISCUSS labels, only the AGREE and the DISAGREE labels play a role when it comes to constructing the bipolar argumentation graphs on which the DF-QuAD-based stance aggregation is performed.

[8] https://nlp.stanford.edu/software/CRF-NER.html

HYPER-PARAMETER	VALUE
Batch size	16
Dropout	0.25
Recurrent dropout	0.25
Units (dimensions of output space)	64

Table 2: Hyper-parameters for training RNN models.

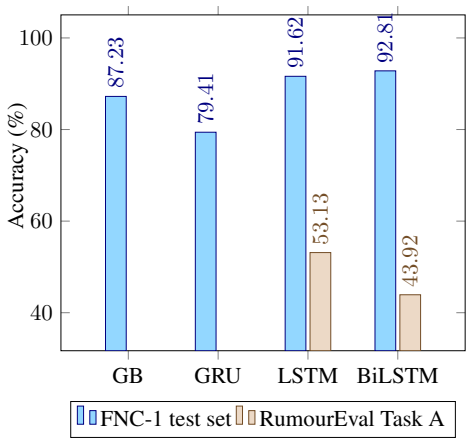


Figure 5: Accuracy of classifiers for 3-way stance problem.

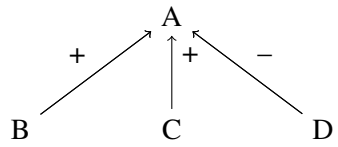
5.2 Stance Aggregation

This section outlines the techniques we employ for aggregating stance labels. Stance aggregation is performed on the RumourEval dataset. We compare the performance of three stance aggregation methods for aggregating both the gold standard stance labels provided for the RumourEval Task B dataset and also the labels generated by the LSTM and BiLSTM models. We choose to only use the predictions generated by the LSTM and BiLSTM models because they display the best test performance on the RumourEval Task A dataset, as we will see in Section 6.

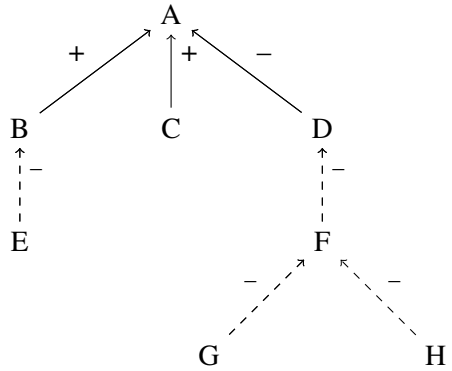
Dataset	Model	AGREE			DISAGREE			DISCUSS		
		P	R	F1	P	R	F1	P	R	F1
FNC-1	GB	**.831**	.736	.781	.570	.322	.412	.926	**.972**	.934
	GRU	.645	.685	.665	.402	.244	.304	.876	.887	.882
	LSTM	.817	**.878**	**.846**	.652	**.493**	.562	**.964**	.955	**.960**
	BiLSTM	.829	.840	.835	**.676**	**.493**	**.570**	.949	.965	.957
RUMOUR EVAL	LSTM	.166	**.490**	.248	**.160**	.0119	.0222	.753	.513	.610
	BiLSTM	**.178**	.430	**.252**	.105	**.0448**	**.0628**	**.759**	**.576**	**.655**

Table 3: Precision (P), recall (R), and F1-score (F1) of stance detection classifiers on FNC-1 test set and RumourEval dataset (see Section 5).

5.2.1 Aggregation via DF-QuAD Semantics



(a) Only direct reply (DR) stance relations for direct and nested replies



(b) Direct reply (DR) and nested reply (NR) stance relations for direct and nested replies.

Figure 6: Examples of constructed BAFs. A is a source tweet from RumourEval. B, C, D, E, F, and G are all replies. Direct attack and support relations are drawn with solid lines. Nested attack and support relations are shown with dashed lines.

Each conversation thread in RumourEval takes a form similar to the example given in Section 4.2. Argument A in Figure 6 is the claim for which we aim to predict the veracity. A is a source tweet (i.e., start of the conversation thread), so it forms the root node of the graphs shown in Figures 6a and 6b. We construct two BAFs: (1) a BAF in which attack and support relations only exist between source tweet, in this case argument A, and direct replies, as dictated by the stance detection classifier, and (2) a BAF with additional relations between reply tweet nodes, accounting for nested replies as well as direct replies. Figure 6 shows that A has three direct replies in the conversation thread; these are B, C, and D. Only these four arguments (A, B, C, D) are present in the flat BAF described in (1) above. The BAF illustrated in Figure 6b incorporates the responses (arguments and counter-arguments) to A's replies B, C, and D. B is attacked by argument E, and D is attacked by F, which is subject to two counter-arguments G and H. The motivation for the latter graph construction, which incorporates both direct and nested reply tweets, is to learn the credibility of replies through their relation to each other, and incorporate this in the aggregation indirectly, via their dialectical strength. This reflects the acceptability of the claim in the context of the arguments formed with texts that support and refute it, as opposed to the credibility used to compute credibility-weighted averages, which is often based on meta-data pertaining to the source of the claim.

6 Results

Here we discuss the results obtained for both stance detection and stance aggregation for veracity prediction. We evaluate the effectiveness of the four classifiers given earlier for stance detection by cross-validation on the FNC-1 dataset, and the choose the two best performing such classifiers on the RumourEval Task A dataset. We then evaluate the effectiveness of methods for predicting the veracity of the rumour claims presented in the RumourEval dataset: these are, in addition to the two DF-QuAD-based methods presented earlier, a standard credibility-weighted average baseline.

6.1 Stance Classification Performance

As expected the stance detection classifiers performed well on the FNC-1 3-class task, but quite poorly on the RumourEval Task A dataset (see Ta-

ble 3). This is most likely because of the paucity of DISAGREE examples in the training data. The LSTM and BiLSTM classifiers recorded the best performance on the FNC-1 test set. For this reason, we chose to use these two models for predicting stance labels on RumourEval Task A.

6.2 Aggregation Performance

Table 4 summarizes our stance aggregation results, from which it can be seen that the DF-QuAD-based aggregation methods exhibit comparable or better performance than the non argumentation-based baseline. Figure 7 shows the accuracy achieved by each method for the gold standard labels and the predicted labels. Further error analysis is given in the confusion matrix for each of the aggregation methods provided in Figure 8.

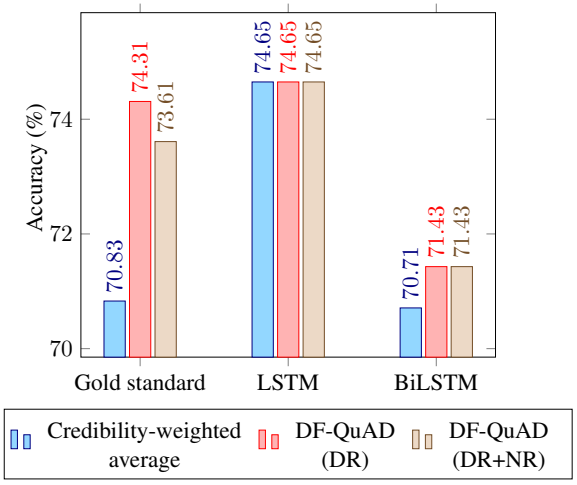


Figure 7: Comparison of stance aggregation accuracy achieved by each method on gold standard labels, and LSTM and BiLSTM stance detection labels.

6.2.1 Baseline

The baseline we devised for our experiments computes a credibility-weighted average of the disagree and agree stance labels relating to a claim.

For the credibility-weighted average we simply defined the credibility to be the number of followers of the account that posts the reply. Since it is often the case that spam accounts will have many followers that are not genuine (i.e., we assigned any account that does not have a profile photo a credibility of zero, assuming that this is not a genuine account. We normalized the Twitter user credibility for each reply in a conversation.

Actual \ Predicted	F	T
F	38.3%	61.7%
T	13.4%	86.6%

(a) Gold standard labels weighted average

Actual \ Predicted	F	T
F	53.2%	46.8%
T	15.5%	84.5%

(b) Gold standard labels DF-QUAD (DR)

Actual \ Predicted	F	T
F	51.1%	48.9%
T	15.5%	84.5%

(c) Gold standard labels DF-QuAD (DR+NR)

Actual \ Predicted	F	T
F	5.0%	95.0%
T	3.0%	97.0%

(d) BiLSTM labels weighted average

Actual \ Predicted	F	T
F	7.5%	92.5%
T	3.0%	97.0%

(e) BiLSTM labels DF-QuAD (DR)

Actual \ Predicted	F	T
F	7.5%	92.5%
T	3.0%	97.0%

(f) BiLSTM labels DF-QuAD (DR+NR)

Actual \ Predicted	F	T
F	7.9%	92.1%
T	1.0%	99.0%

(g) LSTM labels weighted average

Actual \ Predicted	F	T
F	10.5%	89.5%
T	1.9%	98.1%

(h) LSTM labels DF-QuAD (DR)

Actual \ Predicted	F	T
F	10.5%	89.5%
T	1.9%	98.1%

(i) LSTM labels DF-QuAD (DR+NR)

Figure 8: The confusion matrix for each of the aggregation methods performed on the three types of label.

6.2.2 Comparison of Methods

Stance aggregation was performed using four methods, of which two argumentative: one implementation of DF-QuAD on BAFs considering only the argumentation relations on direct reply edges of the BAFs, and another which considers all relations. We performed a DF-QuAD strength evaluation on both the flat and layered BAFs. We interpreted a value of the DF-QuAD strength function (see Section 3) which is > 0.5 to be a true label, otherwise the rumour claim is labelled false.

For all three types of labels, the aggregation-based evaluation either beats the baseline or performs equally as well. Furthermore, the LSTM and BiLSTM predicted labels achieve aggregation accuracy results that are very similar to those achieved using the gold standard labels. The BiL-

	Stance aggregation method	Veracity Assessment (RumourEval Task B)					
		FALSE			TRUE		
		P	R	F1	P	R	F1
Gold standard labels (RumourEval Task A)	CREDIBILITY-WEIGHTED AVERAGE	.581	.383	.462	.743	**.866**	.800
	DF-QUAD (DR)	**.625**	**.532**	**.575**	**.789**	.845	**.816**
	DF-QUAD (DR + NR)	.615	.511	.558	.781	.845	.811
LSTM stance detection labels	CREDIBILITY-WEIGHTED AVERAGE	**.750**	.079	.143	.746	**.990**	**.851**
	DF-QUAD (DR)	.667	**.105**	**.182**	**.750**	.981	.850
	DF-QUAD (DR + NR)	.667	**.105**	**.182**	**.750**	.981	.850
Bidirectional LSTM stance detection labels	CREDIBILITY-WEIGHTED AVERAGE	.400	.050	.089	.719	**.970**	.826
	DF-QuAD (DR)	**.500**	**.075**	**.130**	**.724**	**.970**	**.829**
	DF-QuAD (DR + NR)	**.500**	**.075**	**.130**	**.724**	**.970**	**.829**

Table 4: Precision (P), recall (R), and F1-score (F1) of the stance aggregation methods when applied to both gold standard stance labels and the stance labels predicted by the LSTM and bidirectional LSTM trained stance detection classifiers.

STM labels give the worst performance of the three label types. This is likely related to the fact that, although the BiLSTM classifier outperforms the LSTM classifier on the FNC-1 dataset (see Figure 5), it does not accurately predict RumourEval Task A labels as well as the LSTM – particularly DISAGREE labels. As expected, the gold standard tweet labels show the best performance for the two DF-QuAD aggregation methods. They also show comparable results to the LSTM labels, which however are likely to be unreliable because of the classifiers inability to generalize well.

7 Conclusions and Future Work

We have proposed a method for veracity prediction based on a form of argumentative aggregation rather than credibility-weighted average of stance labels. We used stance label predictions for relation-based argument mining to generate bipolar argumentation frameworks (BAFs). We then evaluated the dialectical strength of arguments in these frameworks as a form of aggregation for veracity prediction. Empirical results on a combination of the FNC-1 dataset for stance detection and RumourEval dataset for veracity prediction show that modelling various stance labels within a bipolar argumentation framework may offer a promising new approach to fake news detection via stance detection and dialectical aggregation.

However, there were a number of limitations in our study, in particular the size of the training data and the unbalanced labels of the training data, resulting in stance detection classifiers that performed poorly on the unseen RumourEval dataset. In order to improve the performance of the classifiers we could incorporate an attention mechanism in our RNN architectures. Furthermore we could train the models on hand-crafted lexical features in addition to word embeddings. In addition, the rumour understanding dataset and the features described in Turenne (2018) could be employed for further experiments into gradual argumentation evaluation of stances.

In order to draw further conclusions about the usefulness of dialectical strength in the task of stance aggregation, studies should be conducted on more robust classifiers. The limitations of the training datasets and classifiers developed from this training data mean that the conclusions we can infer are limited. Also, as we elucidate in Section 3, the nature of the data – conversations taken from social media – also restricts the observations we can draw from our findings. Furthermore, it would be worthwhile to investigate the performance of other gradual semantics for BAFs, as well as non-gradual semantics, to evaluate the strengths of claims in BAFs.

For future work, it would also be worthwhile to explore how BAFs extracted from stance detection classifiers, and the dialectical relations between the arguments in these BAFs, could be used to provide explanations for the veracity prediction of the claim. These explanations would hopefully provide clarification about why a veracity label – true or false – was decided, as well as which evidence or counter-evidence arguments were most pivotal in arriving at that judgement.

References

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58.

Lucas Carstens and Francesca Toni. 2015. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34.

Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005. Gradual valuation for bipolar argumentation frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 366–377. Springer.

Oana Cocarascu and Francesca Toni. 2018. Combining deep learning and argumentative reasoning for the analysis of social media textual content using small data sets. *Computational Linguistics*, 44(4):833–858.

Costanza Conforti, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Towards automatic fake news detection: Cross-level stance detection in news articles. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 40–49.

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-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

William Ferreira and Andreas Vlachos. 2016. 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*, pages 1163–1168.

Bilal Ghanem, Paolo Rosso, and Francisco Rangel. 2018. Stance detection in fake news a combined feature representation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 66–71.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Nayeon Lee, Chien-Sheng Wu, and Pascale Fung. 2018. Improving large-scale fact-checking using decomposable attention models and lexical tagging. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1133–1138, Brussels, Belgium. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

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 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 767–776, New Orleans, Louisiana. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Antonio Rago, Francesca Toni, Marco Aurisicchio, and Pietro Baroni. 2016. Discontinuity-free decision support with quantitative argumentation debates. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016*, pages 63–73. AAAI Press.

Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *arXiv preprint arXiv:1707.03264*.

Jingbo Shang, Jiaming Shen, Tianhang Sun, Xingbang Liu, Anja Gruenheid, Flip Korn, Ádám D Lelkes, Cong Yu, and Jiawei Han. 2018. Investigating rumor news using agreement-aware search. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2117–2125. ACM.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. Proceedings of the first workshop on fact extraction and verification (fever). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*.

Nicolas Turenne. 2018. The rumour spectrum. *PloS one*, 13(1):e0189080.

Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. 2019. Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 836–837. ACM.