Multi-Task Learning Framework for Stance Detection and Veracity Prediction

Fatima Alkhawaldeh $^{a, b, *}$, Tommy Yuan $^{a, c}$ and Dimitar Kazakov $^{a, d}$
[a] Department of Computer Science, University of York, YO10 5GH, York, UK
[b] E-mail: ftma500@york.ac.uk
[c] E-mail: tommy.yuan@york.ac.uk
[d] E-mail: dimitar.kazakov@york.ac.uk

Abstract. As more people rely on online media, it becomes more challenging to identify trustworthy information. As a result of this increased complexity, stance detection and rumour detection have gained prominence. Although both tasks are highly correlated and should be performed concurrently, most existing models train them independently. Additionally, while each target topic may contain numerous conflicting claims, previous work treated each claim independently, resulting in conflict claims wrongly assigned with the same truth label. Because some lengthy rumour posts cover a wide range of topics, determining the positions of the posts can be done with a variety of target topics. Existing models may take a biased position toward the correct target topic or the incorrect target topic, resulting in an incorrect determination of veracity. The purpose of this article is to address these problems by proposing a framework for stance detection and veracity prediction that takes into account source credibility and compares the strength of arguments in order to forecast the truth. Experiments are conducted using two well-known datasets: Emergent and RumourEval-2019. On the gold-standard datasets, the results demonstrate that the proposed framework outperforms other methods.

Keywords: Stance Detection; Rumour Veracity; Truth Discovery; Bipolar Argumentation

1. Introduction

Compared with traditional media, social media has recently emerged as a low-cost and convenient way to spread rumour, increasing incorrect and misleading information. Rapid dissemination of information, particularly in breaking news, fosters the spread of unverified and unchecked information. Because it is difficult and expensive to hire qualified journalists and other experts to verify published posts, an automated process for verifying unverified information in published articles is required to speed up the process. Information can be disseminated in a variety of formats, such as text, video or image. The focus of this paper is to detect and verify text-based rumours.

Unverified information that can later be believed to be true or false is a rumour [1]. According to Harsin [2], rumour is a claim that is debatable in its veracity, ambiguous in its intent, and originates from an unknown source, such as ideological or partisan sources. False rumours in the news domain are articles published by a news source that misled the reader [3]–[5]. When rumours have a low veracity value, they are called fake news [1], [6], [7]. Numerous rumours have been proven to originate from either fake news or Hyperpartisan websites [8]. Liu and Wu [9] used network embedding to create user representations and demonstrated that user credibility is critical for determining rumours' veracity. Zubiaga et al. [10] define four components of a rumour classification system: i) rumour detection, which determines whether a statement is verified (rumour or non-rumour), ii) rumour tracking, which collects feedback on a rumour as it spreads, iii) rumour stance classification, which determines how the general public feels about the rumour's truth, and iv) rumour veracity classification, which determines the rumour's veracity. The purpose of stance detection is to identify the various attitudes, e.g., agree, disagree toward a particular claim. The task for veracity prediction is to determine whether a rumour is true, false, or unverified.

This work addresses three issues identified from the literature that contribute to the failure of veracity prediction systems to achieve acceptable detection performance. The first problem is that stance detection and veracity prediction are separately trained and learned. Even though both stance detection and veracity prediction are positively correlated with joint treatment, current research treats them as distinct tasks, either stance detection [11] or veracity prediction [12]. Because it is not always possible to ground claims in knowledge bases (authoritative sources), particularly for emerging claims, the stances of social media users toward claims can provide indicative clues about their veracity. As a result, the two tasks, stance detection and veracity prediction can be learned concurrently to maximise their utility. The article proposes a novel multi-task learning scheme for simultaneously predicting rumour stance and veracity to enhance the performance of a veracity prediction task by leveraging the related task of stance detection, taking into account the strong correlation between claim veracity and the stances expressed in responsive posts.

The second problem is taking stances on lengthy claims with multiple target topics without focusing exclusively on one target topic that receives more attention in response to other claims (replies). For lengthy claims with multiple target topics commented on by multiple claims (replies), previous models attempted to detect the general stance without considering the primary or the most concerned target topic. As a result, the stance decision may be incorrect. Therefore, it is essential to extract a specific target topic and examine the stances taken toward the claim in light of this targeted topic. The purpose of extracting the primary common target topic in our proposed model is to eliminate irrelevant and noisy information. Each replay's stance toward this claim is narrowly focused (Each replay is associated with a user who commented on the source). As a result, detecting the target topic and then deriving a target-specific based claim from a lengthy claim and selecting pertinent data assists in stance detection, whereas noisy data contributes less. Another goal of target topic extraction is to classify all claims with associated target topics according to their likelihood of being a specific target topic, then analyse and rank each argument to determine the strongest one. Each claim's target topic is extracted independently. As a result, the target topics with the most similar embeddings to the primary target topic is selected for analysis alongside the target topic. Rumours from reliable sources are weighted heavily in the outcome, whereas rumours from unreliable sources are ignored.

The final problem is that when multiple claims on the same target topic originate from multiple sources may conflict, they are analysed independently, whereas they should all be considered during the claim checking process. Current models tend to be restricted to assessing the veracity of claims (rumours), rather than distinguishing conflicting claims on the same target topic, which results in disagreements when various sources are commenting on the same target topic. Consequently, conflicting statements about the same target topic can be labelled identically, which is illogical in light of the fact that each rumour is independently checked for veracity. Thus, a choice must be made between many conflicting facts about an entity. Given that many statements are made in response to the same thing, we hypothesise that only one of the claims, not the disputes, is credible. Each claim in a natural language argument expresses an opinion about a particular target topic; by incorporating argumentations into the context, the claim can be processed simultaneously with similar target topics, preventing the labelling of competing rumours with the same truth value. The strength of an argument for the proposed system is based on various facts and characteristics derived from either the original content and its account credibility or the replayed content and its account credibility. This is remarkably similar to theories of truth-discovery, which argue that truth is discovered by argumentation.

This paper focuses on the combination of the two tasks, namely ATD., in SemEval-2019 Task 7 (Rumour Eval 2019) as a single task, using Twitter and Reddit [13] as well as the Emergent dataset [14], to solve the problem with the lengthy rumour with various target topics. In addition, we consider applying truth discovery to integrate information about source credibility, i.e., it does not account for equal contributions from various sources. The proposed framework incorporates a source credibility metric to compare the strength of arguments in order to forecast the truth, taking into account both the arguments backed up by supporting sources and the claims rebutted by attacking sources.

The contributions of this work are:

- We propose a novel framework for tackling stance classification and veracity checking concurrently. As far as we are aware, this is the first work to employ argumentation-based truth discovery.

- A novel model for the optimal target-specific based claim generator is proposed for the lengthy rumour with multiple target topics while keeping the document's primary target topic in mind. Target topic extraction enables the examination of all pertinent arguments in order to determine the truth by the same target topic.
- The application of the Emergent and SemEval 2019 datasets demonstrates that this framework outperforms existing methods in both stance classification and veracity checking. In addition, this framework can address new rumours (unseen data) that spread rapidly on social media and mitigate their negative consequences.

The remainder of this essay is structured in the following manner. Section 2 reviews relevant literature. Section 3 introduces the proposed Argumentation-based Truth Discovery Model and demonstrates how it can be expressed in bipolar argumentation. Section 4 discusses the experiments and analyses the result, Section 5 concludes the paper.

2. Related Work

We provide a brief overview of related work on rumour detection on social media. Generally, rumour detection methods make use of news content and social contexts [15]. Linguistic-based features are primarily extracted in content-based approaches. Rumour articles contain indicators that can be used to tell the difference between fake and real news. Several attempts have been made to automatically verify the veracity of news articles based on their content, for example, Ciampaglia et al. [16] transform Wikipedia into a knowledge network, where relevant data is arranged in a graph. Conroy et al. [17] considered linguistic information a critical factor, for example, style-based fake news detection, which takes the article's language style into account [18].

Additionally, websites (e.g., Emergent, PolitiFact, and Snopes) independently verify user claims or provide information discovered on the website by human professionals. Along with news content, the social context surrounding news articles contains valuable information that can be used to identify fake news. The characteristics of social context-based approaches are as follows: user-based approaches extract data from user profiles, network-based approaches extract data by creating particular networks [19], and post-based approaches reflect users' social responses in terms of reputation, topics, or stance-based and propagation-based [5].

To detect rumours, various methods, such as classic machine learning, are used. Machine learning algorithms use a set of predefined linguistic features and a large amount of labelled data. Others have discovered that implementing modern neural network models that use pre-trained word vectors and embedded representations improve quality [20], [21]. Zhao et al. [22] employ a cluster ranking algorithm to prioritise tweet clusters based on their likelihood of being rumours. Zubiaga et al. [23] also employ an unsupervised approach, employing a sequential classifier to learn rumour characteristics such as lexical and temporal structure. Certain works, such as Saikh et al. [24], combine textual entailment and stance classification using statistical machine learning and deep learning techniques. We discuss briefly several models that are relevant to the primary objective of the work. It begins with stance classification, progresses to veracity checking, joint stance and veracity prediction, and truth discovery.

2.1. Stance Classification

Ferreira and Vlachos proposed and released the Emergent dataset, which contains 300 PolitiFact-labelled rumours[14] that predicts whether a claim is for, against, or observed based on news headlines' lexical and semantic features. As baselines for the work by Zhang et al. [25], Table 1 illustrates some state-of-the-art models that employ various models for the Emergent dataset's stance detection task. Each model was trained and tested on the same dataset, and their performance was quantified in terms of relative score. The relative score evaluates a model by dividing it into two subtasks: related and unrelated and then categorising the related as agree/disagree/discuss. Their model predicts relatedness first, contributing 25% to the relative score, followed by stance from the three related classes, which contributes 75% to the relative score.

Table 1

A summary of work on stance detection: State-of-the-Art models on the Emergent dataset

The model	**The implementation details**	**Relative score (%)**
LSTM (BiLSTM)	Stance Detection with Bidirectional Conditional Encoding. The encoded claim is used as initial states to encode the evidence [26] after the 100-d GloVe word embedding is applied [27]	81.37
Attentive CNN (AtCNN)	For both claim and evidence feature representations, the convolutional neural network is used and attention mechanism to extract the most relevant features [28]	83.56
Memory Network (MN)	A combination of convolutional and recurrent neural networks by an end-to-end memory network is implemented [11]	85.92
Ranking Model (RM)	Ranking model to maximise the difference between the four stances representation agree, disagree, discuss, unrelated [29]	87.69
Official Baseline (OB)	gradient boosting decision trees model for stances [30]	74.86
Logistic Regression (LR)	After checking whether the source is related or not by n-gram matching and rule-based methods. The stances: agree, disagree and discussed are decided by Logistic Regression [31]	83.45
Gradient Boosted Decision Trees (GBDT)	Apply Gradient Boosted Decision Trees to detect related stance and apply another Gradient Boosted Decision Trees to detect the remaining three stances [32]	87.53
Multi-Layer Perception (MLP).	Cosine similarity between claims and evidence, and Multi-Layer Perception for the four stances [33]	85.43
Hierarchical representation of a neural network	Hierarchical representation of these classes combines agree, disagree, and discuss classes under a new related class where the hierarchical architecture alleviates the class imbalance problem. One neural network layer for related stance detection, and the second layer is for the three stances detection [25]	**89.30**

The model in [25] compares their performance with four-way classification baselines (OB, MLP, BiLSTM, AtCNN, MN and RM) and demonstrates state-of-the-art accuracy performance. They develop a two-layer neural network that learns from this hierarchical representation to alleviate the class imbalance problem; the first layer for relatedness stance detection and the second layer is for the agree, disagree and discuss stances. They study the various levels of dependence assumptions between the two layers: controls the error propagation between the two layers using the Maximum Mean Discrepancy MMD regularizer. They demonstrate that their work outperforms the state-of-the-art accuracy for the stance detection task.

Some observations are made by Zhang et al. [25]: (1) Dependency-based CNN and constituency-based CNN improves overall performance by detecting the claim's complex syntactic structures and target-specific based claim. It can capture long-distance syntactic dependency. (2) Using the Manhattan distance to infer the claim and the target-specific based claim underlying semantic similarity based on the vector representation (final hidden states), help to capture the semantically equivalent of claim and target-specific based claim. (3) The effectiveness of attention mechanism via emphasizing the words necessary to the semantics of the claim and target-specific based claim by automatically search for the most relevant parts of an input sequence and assigns weights to those parts. Our models' performance on Emergent data is compared to the best performing model reported in [25], as shown in the last row of table 1.

The Rumour Eval 2019 competition attracted a diverse field of competitors [13]. The competition's top-performing system for the stance detection task is described by Yang et al. [34], using the Twitter and Reddit datasets. They infer the conversation chain from the source post to the replies using Bi-LSTM and Transformer, relying on features such as the number of question words, the presence of rumour words, false synonyms, and false antonyms. The second-ranked system, Fajcik et al. [35], makes use of an ensemble of BERTs, while the third-ranked system, Baris et al. [36], makes use of a pre-trained representation with OpenAI GPT. Pre-training representation models [37] and ELMO [38] have demonstrated promising results in which each word's representation is based on the context in which it is used.

Table 2

RumorEval 2019 test results for Task A: Stance Detection

System	**MacroF**
Khandelwal’s [39] Method	0.6720
BLCU NLP. [34]	0.6187
BUT-FTT [35]	0.6167
EventAI [40]	0.5776
Hierarchical graph convolutional network GCN-RNN [41]	0.540
Top-down tree structure using a recursive neural network TD-RvNN [12]	0.509
UPV-28-UNITO [36]	0.4895
HLT(HTTSZ) [37]	0.4792

2.2. Veracity Checking

The majority of veracity checking systems have been developed using the FEVER dataset [42]. FEVER is a large-scale dataset for fact extraction and verification that contains 185,445 claims and their supporting evidence. In the first FEVER shared task, the Bi-Directional Attention Flow (BiDAF) network [43] outperforms Neural Semantic Matching Networks (NSMNs) [44] and contextualised representations [45] of a pre-trained BERT [46]. BiDAF [43] generates two vector sequences for claim and evidence, and the attention layer computes the attention scores before sending them to the output layer, which computes the semantic similarity between the original sequences and the new vectors. Finally, the label is generated by the output layer. In NSMNs [44], an LSTM matching layer performs semantic matching on the encoded claim and evidence sequences and sends the output to the label generation's output layer. In BERT, 12 self-aware encoder layers are combined with a classification layer to generate a highly embedded representation of the claim and evidence; the classification layer then uses this representation to generate labels.

The models described by Enayet & El-Beltagy [47] achieves competitive results in the SemEval 2017 rumour detection tasks. The best performing model for the veracity prediction task presented in the Rumour Eval 2019 competition [48] is presented in Li et al. [40], which employs a variety of classifiers (Support Vector Model, Random Forest, Logistic Regression) with features derived from the LSTM attention network. Tables 3 shows the systems that performed best in the rumour detection task in the RumorEval 2019 competition [13].

2.3. Multi-task approach for joint prediction of rumour stance and veracity

Various studies have established that stance detection is the most critical task for rumour verification. [22], [41], [49]–[55]. For example, Khandelwal [39] proposed a framework for jointly predicting rumour stance and veracity using multi-task learning. Recently, the multi-task approach for joint prediction of rumour stance and veracity using deep learning models such as BiLSTM [39] outperforms previous methods on both the SemEval 2019 Task 7 dataset for rumour stance classification and veracity prediction. Khandelwal [39] obtains the post representation using sliding window-based self-attention and a pre-trained Longformer [56]. As a result, we rely on Khandelwal [39] to assess the model's performance on the SemEval 2019 Task 7 dataset.

Certain studies use stance detection and the labels extracted from them as an input feature to improve the performance of veracity prediction models, which are significant indicators for predicting the veracity of rumours [49], [22], [47], [57]–[59], [41]. They combine stance detection and rumour veracity classification tasks by utilising various forms of multi-task learning, including parallel feature learning [57]–[59], [7], and hierarchical design [41]. Ma et al. employ a GRU layer for each task in [58], and the tasks share a GRU layer to obtain patterns common to both tasks. Joint learning is used in Wei et al. [41], similarly to Ma et al. [58]. Both models are composed of a common layer and task-specific layers; neither model incorporates user information, whereas Li et al. [7] combine the attention mechanism with user credibility information. Tables 2 and 3 compare the performance of various methods on the SemEval dataset for classification of rumour stances (single task) and veracity (multi-task).

Table 3

RumorEval 2019 test results for Task B: veracity prediction.

System	The implementation details	MacroF
Li et al.'s model [7]	Incorporated features relating to user credibility in addition to other tweet-related features.	0.606
Khandelwal's [39] Method – Top N_s using (D + E + F)	This method is based on multiturn conversational modelling using a transformer-based model, natural language processing feature extraction from conversations, collaborative rumour stance learning, and veracity classification.	0.5868
Hierarchical- predicting rumour Stance and Veracity PSV [41]	Utilises a graph convolutional network to model the relationships between each thread's posts.	0.588
EventAI [40]	The authors used the cosine similarity of tweet replies to the source message as a consistency metric. The consistency of the tweet is regarded as a valuable characteristic for rumour detection.	0.5765
MTL2 (Veracity+Stance) [57]	A method for performing multiple tasks without the use of a task-specific layer	0.558
BranchLSTM+NileTMRG [57]	Utilises the same characteristics as the stance classification system but generates a single output for each branch. The thread's veracity prediction is then determined by majority voting over per-branch outcomes. NileTMRG (is a linear SVM that employs a bag-of-words representation of the source tweet, concatenated features defined by the presence of URL, hashtag, and the proportion of supporting, denying, and querying tweets in the thread.	0.539

2.4. *Truth Discovery*

Rumour verification is challenging because relevant sources contradict published posts. Several truth discoveries for estimating the reliability of attacking/supporting sources have been proposed to select the most trustworthy ones to address this issue. The purpose of truth discovery algorithms is to resolve inconsistencies in information obtained from multiple sources and on the same subject [60]. Due to the variety of qualities and reliabilities of sources attacking/supporting a claim that should be considered when determining the veracity of information, analysing and assessing the information's credibility based on a single source often fails because it may be biased and is, therefore, less trustworthy than multiple supporting sources.

Algorithms for truth discovery have been expanded from data mining and crowdsourcing perspectives to other perspectives such as argumentation which may play a significant role, as discussed in Singleton's papers [61], [62]. There are contradictions between information about a particular object in both truth discovery and argumentation issues: using conflicting facts to discover the truth and attacking arguments to make a claim. Singleton [61], [62] continues by stating that truth discovery can be mapped to a particular type of argumentation, such as bipolar argumentation, in which opposing arguments support and attack one another, to identify 'acceptable' arguments from a collection of conflicting arguments.

In general, there are four types of methods for truth discovery that have been used in previous research.:

i. Iterative methods where the trustworthiness of sources and the confidence of claims from each other are computed iteratively and until convergence [63],
ii. The optimisation that measures the difference between the information provided by sources and the truth-based methods [64],
iii. Probabilistic graphical model-based methods where expectation maximisation is commonly used to infer the latent variables (parameters of truth and source reliability) [65],
iv. Neural network [66].

TruthFinder [67] and Voting [68] are two methods for iteratively updating source reliability and facts from multiple conflicting data sources. Other works [69]–[73] employ a variety of methods to ascertain the truth, including information extraction techniques such as entity profiling in [71] and knowledge graphs in [73]. Recently, [74]–[76] proposed a framework for truth discovery as an optimization problem, in which truths and source reliability are iteratively updated.

Other works employ probabilistic approaches. Source reliability is treated as a random variable in the probabilistic models, and the likelihood or posterior distributions of multi-source data are maximised, as Wang et al. [77] demonstrated by developing a maximum likelihood estimator for source reliability. Samadi et al. [78] used the Probabilistic Soft

Logic (PSL) framework to estimate source reliability and claim correctness, whereas in Nakashole & Mitchell [79], language objectivity analysis is used to determine the veracity of value in addition to Subject-Predicate-Object (SPO) triplets. Other individuals employ various techniques to ascertain the truth; for example, Wang et al. [80] categorise sources and values according to their intended users and then assess the information's credibility. Additionally, probabilistic graphical models with three measures of silent, false spoken, and true spoken rates can be used [81], as well as generative modelling processes as Zhao et al. [82] described. Bayesian analysis can be used to determine the source dependence of data [68]. By considering the confidence interval of the estimation [74], Bayesian probabilistic modelling on the dependencies between source quality, truth, and claimed values [83] estimates source reliability.

In the truth discovery task, neural network models achieve comparable accuracy [84]–[86][66]. Despite showing significant improvements in their rumour detection model when stance information is used, they continue to rely on hand-crafted user features such as follower count and post count to reflect user credibility, which is distinct from stance labels for predicting rumour veracity [49][47][7].

2.5. Analysis

Although these tasks are closely related and that multiple people's stances can be used to predict the claim's absolute veracity, state-of-the-art methods for false information detection are typically proposed for either stance detection veracity checking separately; stance aggregation features are required for effective veracity prediction. Most systems are better at detecting stances or predicting rumour veracity, but not both, according to the RumorEval 2019. This constraint limits the generalizability of models. Additionally, as mentioned in the introduction section, previous works were limited in verifying the veracity of individual claims without considering all claims that discussed the same target topic, which meant that many conflicting claims could be labelled as the same.

Based on the foregoing considerations, this work proposes combining the two tasks and learning them together to aid stance detection-based veracity prediction. Furthermore, to distinguish fake from genuine information, the source reliability of those who first disseminated information about the claim must be checked, so this work incorporates user reliability estimation.

The method used in this work predicts both stance and veracity concurrently and establishes a connection between bipolar argumentation and truth discovery techniques. Unlike the models reviewed above, our proposed method (described in Section 3) concludes an article for a specific target topic (the subject of discussion) and learns representative features of stance detection using a different model architecture. Additionally, our goal is to investigate stance classification as a precursor to automatically determining the veracity of a rumour via joint learning to significantly improve these tasks' performance: stance detection and veracity checking.

3. The Proposed Argumentation-based Truth Discovery Model

3.1. Overview

Previous research reveals that how people react to rumours can help determine their veracity[39] . The success of multi-task learning in stance detection and rumour verification, e.g., by Kochkina et al. [57] and Ma et al. [58], and the observation that people's positions are closely linked to the veracity of the information, prompted us to conduct this study. In contrast to these systems, a new perspective is proposed based on argumentation to consider user trust. To our knowledge, this is the first time that argumentation has been used to model conversations to tackle rumour stance classification and veracity prediction concurrently, to avoid labelling contradictory arguments under the same target topic with the same label. For instance, if both arguments A and B (see below) have the same number of supported (i.e., 3) and refuted (i.e., 2) claims, they are more likely to have the same label, e.g., true claim, despite their conflict.

Argument A: Animal research is the only way to progress at times.
Perspective 1, with support stance: Animal research is only used where other research methods are not suitable.
Perspective 2, with refute stance: Medical breakthroughs can be achieved without doing any scientific or commercial experiment on animals.
Perspective 3, with support stance: Sometimes we have no other choices for Animal research, but then to do some animal testing.
Perspective 4, with support stance: Without animal research, we would have fewer products.

Perspective 5, with refute stance: Animal testing doesn't ensure good results.

Argument B: Animals have a right to live their lives in peace without human interference.
Perspective 1, with support stance: Animal testing significantly harms the animal used.
Perspective 2, with refute stance: Human's rights are a more important consideration than animal rights.
Perspective 3, with refute stance: Innovation often requires the use of animal research.
Perspective 4, with support stance: Medical breakthroughs can be achieved without doing any scientific or commercial experiment on animals.
Perspective 5, with support stance: Animal testing helps humans.

As a result, the following solution was envisaged: simultaneously considering all arguments relating to the same subject. We propose the Argumentation-Based Truth Discovery Model, abbreviated as ATD. The input is in the form of a claim accompanied by a subset of tweet replies, each with a distinct stance: for or against. While stance classification entails per-tweet predictions, verification tasks require only a single output for the initial claim. As in the Emergent dataset, the article's claim source may be longer than the user replies (the related he is supporting and opposing perspectives), and it may also cover a wider range of target topics while the user is only interested in one. For instance, several example candidate target topics from the Emergent dataset article are listed in Table 4. The proposed model is intended to extract the user preferred target topics and generate the primary target topic that will be used to generate a more effective target-specific claim.

Following the primary target topic extraction, the most relevant clauses for the claim are extracted, with only informative information on the target topic. The sequence-to-sequence generator receives the selected clauses and directs the generation process toward the target topic. Numerous Evaluators are used to guiding this model's adversarial training using training signals to optimise its parameters, i.e., determining the difference between generated and ground truth target-specific based claims. Finally, before predicting the claim's veracity, the generated target-specific based claims are used to verify the original claim's stance on the responses claims.

Numerous multi-task neural network models, such as hard parameter sharing networks [43] and soft parameter sharing networks [87]. This paper adopts a soft parameter sharing network model because every task has its own network. A gate mechanism will ensure that only beneficial features of auxiliary tasks are shared with the primary task [87]. Filtering feature flows between tasks is accomplished by assigning them a higher weight (learnable parameter) via a gate mechanism that utilises both sigmoid and scalar weights. The gate mechanism produces a vector of elements in the range [0, 1] that can be used to select (or retain only a subset of) the advantageous features required to perform the given task.

Following the embedding layer, a vector is typically used to represent each word in the input. Our model assigns a private sub-model and a private encoder to each task to extract shared and private features from multiple tasks. We begin by calculating the common representation [h1,..., ht] by encoding the tasks' input embeddings with an encoder such as BiGRU. Then, we employ the attention mechanism to selectively retrieve task-specific information and incorporate gates for useful features that transfer between tasks. For each task, both private and shared features are concatenated.

As with encoders, our models incorporate gates to facilitate the transfer of features between sub-models. A gate g is added to task j when it borrows features from task k to select the most useful ones. The gate g is calculated from the previous layer as Eq. (1):

$$g_{jk}^{l} = \sigma\left(W_{jk}^{l} \cdot F_{k}^{l} + b_{jk}^{l}\right) \qquad (1)$$

Where l means the level of the layers and σ denotes the nonlinear activation of the sigmoid. The output F of gates from task j is calculated by fusing the lower layers Fl from all the tasks together, Eq. (2):

$$F_{j}^{l+1} = \sum_{k\in C, k\neq j} g_{jk}^{l} \odot F_{k}^{l} + F_{j}^{l} \qquad (2)$$

We introduce a task-specific query vector q(k) to calculate the attention distribution α(k) over all positions as in Eq. (3).

$$a_{t}^{(k)} = softmax\left(q^{(k)^{T}} h_{t}\right) \qquad (3)$$

Where the task-specific query vector q(k) is a learned parameter. a task-specific query vector will be used to focus target for conclusion generator and claim for stance detection and rumour veracity

The final task-specific representation c(k) is summarized in Eq. (4).

$$c^{(k)} = \sum_{t=1}^{T} a_{t}^{(k)} h_{t} \qquad (4)$$

Table 4

An example of our proposed model-ATD on emergent data [14].

Initial Source	law360.com
The article	Wonder how long a Quarter Pounder with cheese can last? Two Australians say they bought a few McDonald's burgers for friends back in 1995 when they were teens, and one of the friends never showed up. So, the kid's burger went uneaten and stayed that way, Australia's News Network reports. "We are pretty sure it is the oldest burger in the world," says one of the men, Casey Dean. Holding onto the burger for their friend "started as a joke," he adds, but "the months became years and now, 20 years later, it looks the same as it did the day we bought it, perfectly preserved in its original wrapping." Dean and his burger-buying mate, Eduard Nitz, even took the burger on the Australian TV show The Project last night and "showed off the mould-free specimen" News 9 reports. The pair offered to take a bite of it for charity but were dissuaded by the show's hosts. They have also started a Facebook page for the burger called "Can This 20-Year-Old Burger Get More Likes Than Kanye West?" with more than 4,044 likes as of this writing. Furthermore, they are selling an iTunes song, "Free the Burger," for $1.69, and giving proceeds to the charity Beyond Blue, which helps Australians battle anxiety and depression. (A few years ago, a man sold a 20-year-old bottle of McDonald's McJordan sauce for $10,000. Here's why Mickey D's food seemingly, never decays.)."
Candidate target topics	Australia, Food, Hamburger, McDonald's, Quarter + Pounder …. etc
Extracted primary target topic	McDonald's
Clause Selection	Wonder how long a Quarter Pounder with cheese can last? Two Australians say they bought a few McDonald's burgers for friends A man sold a 20-year-old bottle of McDonald's McJordan sauce for $10,000.
Generated target-specific based claim (original claim to be verified)	For 20 years, two Australian men held a McDonald's Quarter Pounder with Cheese
Stance Detection from different sources	**Source-1**: 9news.com.au **Headline**: Two blokes dared to eat a 20-year-old burger for charity **Stance**: for **Source-2**: mirror.co.uk **Headline**: Is this the world's oldest burger? Man claims to have kept McDonald's Quarter Pounder for 20 YEARS **Stance**: for **Source-3**: examiner.com **Headline**: 20-year-old burger: McDonald's Quarter Pounder looks nearly new after 2 decades **Stance**: observing **Source-4**: techinsider.net **Headline**: 20-Year-Old Quarter Pounder Looks About the Same **Stance**: observing
Overall Veracity Prediction via ATD:	true

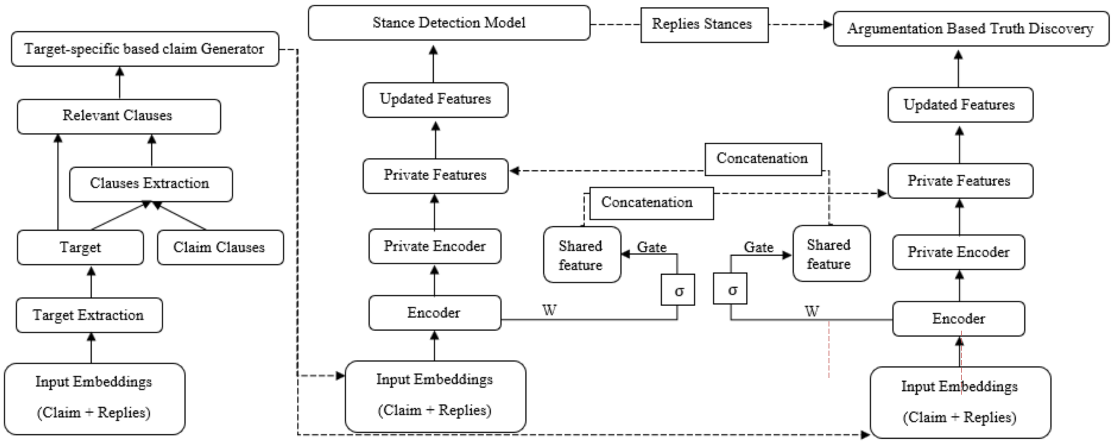

Fig. 1. The architecture of Argumentation-Based Truth Discovery Model.

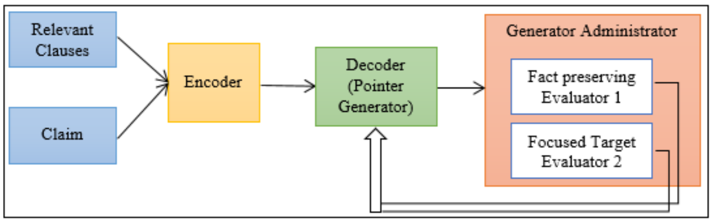
Relevant Clauses
Claim
Encoder
Decoder (Pointer Generator)
Generator Administrator
Fact preserving Evaluator 1
Focused Target Evaluator 2

Fig. 2. Overview of target-specific based claim generator.

The architecture of ATD is shown in Figure 1, with the main components as follows:

- The target-specific based claim generator component employs a seq2seq architecture with attention and copy mechanisms to generate claims focused on a single target topic.
- The stance detection component was developed to determine the position of the article claim relative to other replies.
- The prediction component is created by using Argumentation-Based Truth Discovery to determine whether the claim is valid.

Each component's design is discussed individually below. The example case in table 4 demonstrates the proposed ATD in action.

3.2. *Target-specific based claim generator*

Target-specific based claim generator is a model that conveys a specific stance toward a specific target topic as a key to understanding an argument from its claim, especially if it is a long one. Figure 2 depicts a general overview of this component. Abstractive text summarization is the closest work to this model's target-specific based claim generator; most of them generate summaries by the decoder based on encoded information from the encoder; some use a copy machine to solve the out-of-vocabulary problem [88], [89]. Different earlier approaches [90]–[93] are proposed to capture the central target topic then summarise based on the main target topic. Chen et al. [93] proposed target topic aware summarisation by rewriting the most silent sentences, which achieves the best performance on CNN/Daily Mail benchmark Dataset [94].

This work employs a pointer generator architecture with attention and copy mechanisms to create a claim-target topic-based generator. The pointer generator acts as a decoder, with the selected clauses vectors concatenated with the primary target topic passed to the article encoder serving as the model's inputs. The generator receives the representation outputs from each encoder (decoder). Both encoders and decoders employ a Recurrent Neural Network, namely Bi-GRU encoders and GRU decoders.

3.3. *Extraction of the target topic*

Because a strong target-specific based claim should include the primary objective of the claim, we argue that deducing a claimed target topic is a critical step in conducting lengthy based claims and that this targeted topic should be related to the replay's target topics. The extracted target topic is used to generate the target-specific based claim of an argument based on its article, in which all associated replies adopt a single target topic position. The primary target topic is used to demonstrate or indicate the subject to which the author wishes to direct readers, whereas each claim, particularly the longer ones, may cover a variety of topics or convey the same event via a variety of target topics. As a result, a long claim has generated claim should be focused on the primary target topic. Additionally, the target-specific based claim aids in the detection of stances associated with the claim from replies.

As shown in Figure 3, the purpose of this component is to extract the primary target topic shared by a claim and its associated replies from among candidate target topics. The nouns and replies nouns in the claim must first be extracted. Each noun must be represented as a vector that includes a probability distribution. The Jensen-Shannon Divergence and a distance score greater than or equal to the threshold, set empirically at 0.75, are then used to identify candidate target topics. The Shannon-Jensen Distance (SJD) is an asymmetric version of the Kullback-Leibler Divergence, which uses the difference measure to compute probability distributions [95] which provides a measure of the distance between two probability distributions [96]. It is used to determine text-similarity [97], such as those represented by the p and q vectors in Eq. (5). The relevance score is calculated based on the distance score:

$$1/2(\mathrm{D}(\mathrm{p}||\mathrm{m})+\mathrm{D}(\mathrm{q}||\mathrm{m})) \tag{5}$$

where m = 1/2 (p + q). The two distributions below represent the candidate target topics :
p = as array ([0.10, 0.40, 0.50])
q = as array ([0.80, 0.15, 0.05])
Jensen-Shannon divergence (P || Q): 0.42
Jensen-Shannon distance (P || Q): 0.648, distance is sqrt of divergence.

The JSD [98] explains the contribution of the word I. The smallest divergence indicates that the claims and their replies have a common target topic, Eqs (6):

$$D_{JS,i}(P||Q) = -m_i log_2 m_i + \pi_1 p_i log_2 p_i + \pi_2 q_i log_2 q_i \qquad (6)$$

Where m_i denotes the likelihood that the word I will appear in M.

By determining which of p_i or q_i is greater, we can attribute the contribution to the divergence from the word I to text P or Q. The probabilities of seeing the word I in P and Q are p_i and q_i, respectively.

The Jensen-Shannon Divergence was used to select the candidate target'= topics. The maximum alignment score embeddings of nouns are used to record candidate target topics for claim and replies to link the claimed target topic to the argument's replay target topics. The selected primary target topic has a higher chance of being discussed in the claim and its replies, as explained below. Gao et al. [99] used a max operation over the alignment to select the highly focused noun in the claim by its associated replies , as in Eqs. (7) and (8). This work determines a claim's semantic word alignment based on its embeddings in the replies to model the claim concentrated target topic. Thus, the alignment score indicates the degree to which the word in a claim is targeted at the replies., where $e(A_i^s)^T$ is word embedding in the claim article, and $e(A_{j,n}^c)$ is word embedding in the replay, $target_{i,j,n}$ is the attention for the i-th claim word with the j-th replay word, s is a replay, c is claim, n is article number, i is index word of the replay, and j is the index word in the claim.

$$\mathrm{TARGET\ TOPIC}_{i,j,n} = e(A_i^c)^T e(A_{j,n}^s), \qquad (7)$$

$$maximum_{i,j} = max\left(\left\{target_{i,j,1}, \dots, target_{i,j,T_j^c}\right\}\right), \qquad (8)$$

3.4. *Clause Selection Model*

Clause selection model selects target topic-relevant clauses and eliminates irrelevant and noisy clauses, as our target-specific based claim generator attempts to focus on a single target topic against which other replayed stances can be compared. Clause Picking Module's job is to break down a sentence into clauses and incorporate knowledge of and text information at the clause level. To our knowledge, we are the first to address stance detection by incorporating the critical clause for predicting stances while considering the clauses that correspond to the specific target topic.

Following the first module's selection, the clause selection module selects several clauses about the main target topic, discarding irrelevant and noisy clauses mentioning other target topics, and then generates the claim. When noisy clauses for other target topics are provided and taken into account, the model may make mistakes. The goal of this component is to retrieve the most target topic-relevant clauses while ignoring the rest.

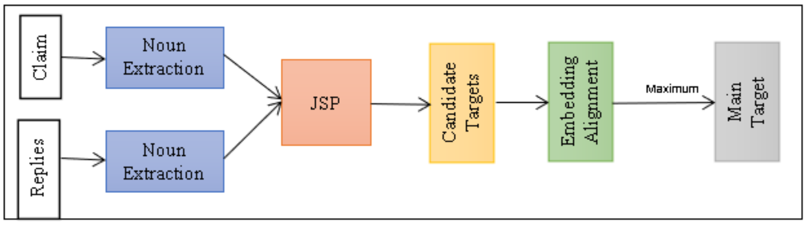

Fig. 3. The general architecture for claim target topic extraction.

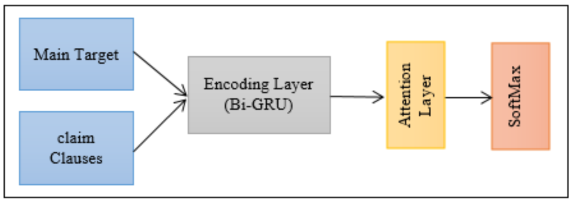

Fig. 4. The Architecture of Clause Selection Model.

This component is made up of two layers. The encoding and attention layers are depicted in Figure 4, which employs bi-directional GRU to capture context clause representations for each clause relevant to the main target topic by concatenating the target topic and each clause cl in the claim. To learn the hidden semantics of words, this model employs GRU, which is more efficient than LSTM training [97]. This module employs two GRU neural networks: a forward GRU and a backward GRU, which process the sentence from left to right and reverse order, respectively, and handle the word vectors in order. Finally, the forward-GRU and backward-GRU units are concatenated to learn the claim's bidirectional semantics and each clause in the article to emphasise the claim's importance. Then, the attention mechanism is used to capture the valuable information the article clauses. The encoding layer does this for both clauses and target topic (h_i for target topic h_j for clauses).

The attention layer produces clause vectors by focusing on words that are relevant to the target topic context. Attention methods would focus on the terms in the clause that are concatenated with the target topic and combine their representations to form a clause vector. The clause representation of the target topic is indicated by clr. To produce contextual information based on the target topic representation, the attention mechanism decides the weight assigned to each target topic-clause representation. The weight of each clause in relation to its representation of a specific target topic will be determined using Eqs. (9) – (12):

$$a_{vi} = avg\left(\left\{h_{i,1}^{c}, h_{i,2}^{c}, \dots, h_{i,T_i^c}^{c}\right\}\right), \tag{9}$$

$$m_i = tanh\left(W_{cl} \cdot \left[a_{vi}; h_{j,T_j^{cl}}^{cl}\right] + b_{cl}\right) \tag{10}$$

$$a_i = softmax(m_i) = \frac{exp(m_i)}{\sum_{t=1}^{Cl} exp(m_t)} \tag{11}$$

$$clr = \sum_{i=1}^{|cl|} a_i \,.\, h_j^{cl} \cdot \tag{12}$$

$h_{j,T_j^c}^{cl}l$ is the state of the clause, cl is clause and c is a target topic, a_{vi} is the average of hidden states for the target topic, a_i is attention weights, and the clause representation clr is calculated based on the attention vectors a_i. In this model, to select relevant clauses for the target topic, conditional probability using SoftMax Layer is used to perform the target topic clause's relevant classification. Then, feeding the clause representation clr to a SoftMax classifier. This model trained by cross-entropy, W and b are the parameters for the model. W is the weight matrix, and b is the bias. The final output o is obtained by Eq. (13).

$$o = W^{*}\, clr + b \tag{13}$$

Loss function computed by cosine similarity between target topic embedding and hidden state of the t-th clause. The similarity is the relatedness between each word annotation hij and the Target' topic representation.

3.4.1. Article (Relevant Clauses) and Claim Encoder

A bidirectional GRU is used to get both the context before and after the word. first, the Word Embeddings method implemented in this work is Glove [27] and word2vec [100], which work better. The encoder generates a state for each input word by BiGRU for the target topic and claims to obtain the context representation around a word. The claim includes all relative clauses retrieved by the clause selection model. $GRU_{\overrightarrow{evi}}\ and\ GRU_{evi^{\leftarrow}}$ are the forward and backwards representation, h_i denotes hidden state. $\left[\vec{h}_i, h^{\leftarrow}{}_i\right]$ is the merge of the forward and backward hidden state, *evi* is the claim and *c* is the target topic. h_i and h_j are the annotations for claim and target topic. It used to compute a hidden representation for claim from both directions and the same for claim encoding.

3.4.2. Decoder

The decoder employs unidirectional GRU, with each decoder time step receiving claim concatenated with its Target' topic representation as input. The decoder begins the decoding process to generate the target-specific based claim from the claim based on the input encoder's final state and target topic representation. At each decoder time step, the target topic embedding is fed as input to allow the decoder to change the output sequence and generate a statement about the primary Target' topic.

The word distribution is calculated, and the word with the highest probability on the decoder state and context vector output is chosen using the SoftMax function. At each decode time step, a sigmoid activation function is used to choose between two options: copy from the original input or generate from the vocabulary, so is the final article encoder state, Ct is the context vector at time step t from the attention mechanism, and Yt-1 is the predicted output word at time step -1. The attention mechanism identifies the input's relevant parts by learning the decoder to focus on different portions of the claim and target topic at different time steps [101]. This could be accomplished using Eqs. (14) – (16), which were inspired by Nema et al. [102], modified to conform to the proposed model. The attention mechanism for the evidence is applied to help the decoder output focused claim tokens at each step using Eqs. (14) and (15) whereα_{ti} represents weights to each in the claim at each decoder timestep, St is the current state of the decoder at time step t. The final claim representation at time step t is computed in Eq. (16):

Attention mechanism for the claim which assigns weights to each word in the claim at each decoder timestep, h_j^c is claim word hidden states. The final claim representation at time step *t*, which is computed as:

$$a_{t,j}^c = v_{cl} \cdot tanh\left(W_{cl}s_t + U_{cl}h_j^c\right) \qquad (14)$$

$$\alpha_{tj}^c = \frac{exp\left(a_{t,j}^c\right)}{\sum_{j=1}^{|cl|} exp\left(a_{t,j}^c\right)} \qquad (15)$$

$$c_t = \sum_{i=1}^{|c|} \alpha_{t,i}^c \, h_j^c \cdot \qquad (16)$$

$h_i$hidden representation for each time-step for evidence word I, $E_{(Y_{t-1})}$ is the previous word embeddings, st is the current state of the decoder at time step t, and cct is the final claim representation at a time step. The hidden state of the decoder st at each time t is computed as follow, considering the previous state s_{t-1}, the embedding distribution of the claimed Target' topic $target$, the previous claim context vector c_{t-1} and the previous word embeddings, the state is defined as Eq. (17):

$$s_t = GRU_{dec}\left(s_{t-1}, \left[c_{t-1}, target, E_{(Y_{t-1})}\right]\right) \qquad (17)$$

The probability distribution over the output vocabulary o_t, as Eq. (18) to decide the word which has the highest probability is computed from the context vector ct, the decoder state st as follow:

$$o_t = W_g^{(2)}\left(W_g^{(1)}[s_t, c_t] + b_g^{(1)}\right) + b_g^{(2)} (18)$$

Inspired by Hasselqvist et al. [103], The decoder in this work uses the pointer mechanism to decide whether to copy from the original document or generate the vocabulary based on the pointer output; the next word can then be chosen. $p_t^{pointer}$ is used as a switch to select between (a) copying words from the source text via pointing (copying a word from the input sequence by selection according to the attention distribution) or (b) generating a word from the vocabulary by selecting based on Pv in Eq. (19).

$$p_t^{pointer} = sigmoid\left(v_{ptr}^T\left[s_t, E_{(Y_{t-1})}, c_t\right] + b_{pointer}\right) \qquad (19)$$

The generation probabilityp_{tj}^{gen}∈ [0,1] for timestep t is computed as Eq. (20). If p_{tj}^{gen} > 0.5, word is copied from the input determined by the attention distribution where the attention is the highest, else the generator output is used. The probability of generating timestamp t is set to 0. 5 empirically.

$$p_{tj}^{gen} = \frac{exp(o_{tj})}{\sum_k exp(o_{tk})} \qquad (20)$$

The model then generates distribution Pv over vocabulary. It concatenates on Evaluator 1, and 2, (details below) and the decoder's output to guide the decoder. P_v is probability distribution over all words in the vocabulary and gives us the final distribution to expect words. It concatenates Evaluators $evalt_t; evalt2_{t;}$ and the output of decoder st as the input of the output projection layer. The goal of, $evalt_t$and$evalt2_{t;}$is to keep track of the difference between the generated target-specific based claim and the focused Target' topic and the fact that in the next subsection, it will show the details of these variables in Eq. (21).

$$P_v = softmax\left(W_v\left[s_t; E_{(Y_{t-1})}, c_t; evalt_t; evalt2_{t;}\right] + b_v\right)(21)$$

This model uses ROUGE scores to evaluate the generated target-specific based claim's quality [104]. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation used to determine the quality of a summary including ROUGE-1, ROUGE-2 and ROUGE-L, demonstrating the effectiveness of discriminator.

- ROUGE-1: the unigram overlaps describe the overlap of each word between the candidate and reference summaries.
- ROUGE-2: bigram-overlap between the reference summary and the summary to be assessed.
- ROUGE-L: the longest common subsequence between the reference summary and the summary to be assessed.

3.4.3. Generators' Evaluators

This work employs two evaluators allowing the discriminator to provide additional information, denoted as Evaluator 1 and 2, to reduce the noisy and irrelevant generated words by the decoder and guide the generator to focus on the information related to the target topic claim and preserve the fact information. Two Evaluator modules guide the generator through the decoding process, preserving the fact while focusing on the claim's primary target topic. Another advantage of target topic extraction is that it can determine which rumour should be believed among several conflicting claims, rather than comparing the veracity of rumours to their replay stances individually, as other experiments do. The discriminator is a binary classifier that uses a convolutional feature extractor and a sigmoid classification layer to signal the generator. The two Evaluators' details are explained in turn below.

Evaluator 1 to check decoder focused target topic

Evaluator 1 is a discriminator that extracts features from a convolutional neural network and then compares the decoder-focused target topic (target-specific based claim) to the claim-focused target topic. It quantifies the semantic difference between the decoder-focused target topic and the claimed target topic; for example, the claimed target topic is "McDonald's Quarter," but the generated target-specific based claim may emphasise the "charity" target topic.

This model makes use of element-wise difference to simulate the difference between target topic and target-specific based claim attention and then uses the decoder to identify the unfocused target topic. The decoder then uses the difference between the attention distribution and the weighted sum of the document states as the context vector to assist it in producing a target-specific based claim that is more focused on the target topic.

To persuade the generator to focus on the claimed target topic, this model employs a CNN-based discriminator to represent the difference between the generator- and claim-focused target topics. After concatenating the target topic, the sentence vector is generated using the BiGRU word-level encoding module. The model establishes a Bi-directional GRU (Bi-GRU) network taking the sentence representation as input further to study the interactions and information exchanges between sentences. This architecture allows information to flow back and forth to generate new sentence representation. The attention-based CNN model for this evaluator will be used. The target-specific based claim's final target topic representation is fed into an output layer to predict the probability distribution on the target topic is defined as Adm1 via Eqs. (22) – (30). It is trained via cross-entropy minimisation for training target topic-based target-specific based claim generation.

$$target = \frac{1}{m}\sum_{i=1}^{m} e_{x_i} \qquad (22)$$

The attention vector decides which semantic features in each hidden state are meaningful specifically towards the Target' topic, which is calculated through a gated structure, as follows: The focused Target' topic for the original claim

$$\mathrm{score_i} = \tanh\left(\mathrm{h_i^T W_1 target\ topic}\right) \qquad (23)$$

$$\mathrm{\alpha ttention_i} = \frac{\exp(\mathrm{score_i})}{\sum_{j=1}^{n}\exp(\mathrm{score_j})} \qquad (24)$$

we sum up the all attention distributions {αt , αt−1 , . . . , αt−n+1 } and result in νt1

$$\mathrm{vt1} = 1/\mathrm{n}\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{\alpha ttention_i} \qquad (25)$$

$$\mathrm{St1} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{vt1_{,i}\, h_i} \qquad (26)$$

$St1$ represents the focused Target' topic for the original claim

The focused Target' topic for the generated target-specific based claim.

$$score_i = tanh(h_i^T W_1 st) \qquad (27)$$

$$\alpha ttention_i = \frac{exp(score_i)}{\sum_{j=1}^{n} exp(score_j)} \qquad (28)$$

We sum up the latest k attention distributions {α_t , α_{t-1} , . . . , α_{t-k+1} } and result in νt2

$$vt2 = \sum_{i=1}^{k} \alpha ttention_i \qquad (29)$$

$$St2 = \sum_{i=1}^{n} vt_{,i}\, h_i \qquad (30)$$

$St2$ represents the focused target topic by the latest k decoding steps, a.k.a., decoder focused target topic.

The score is a content-based function that encapsulates the semantic relationship between the decoder output target topic and the claimed target topic used to determine each word's relative importance in the target-specific based claim and claim. St is used to encode the relevant information n from the target-specific based claim and the claim. The target topic is represented by the target topic's averaged word embedding.

In order to model the gap between claim-focused target topic and decoder focused target topic, we subtract the claim attention by decoder attention resulting in the attention difference shown in Eq. (31). Then, we use the attention difference to sum up the document hidden states h, Eq. (32).

$$diff = vt1 - vt2 \qquad (31)$$

$$evalt_t = \sum_{i=1}^{n} diff_{,i}\, h_i \qquad (32)$$

Evaluator 2 to check the fact preserving

This model applies a denoising autoencoder to evaluate that preservation is related to the target-specific based claim's fact concerning the claim source, i.e., integrating knowledge from the source article; this is represented as a factual score. After extracting the fact related to the source article's target topic, it applies BiGRU to extract hidden state for both facts in the article and the generated fact. At each decoding time step t, GRU reads the previous output yt−1 and context vector ct−1 as inputs to compute the new hidden state st. Then the context vectors are computed. The fact vector Eqs. (33) – (35) are applied, and Eqs. (36) – (38) for a generated fact. Besides the decoder's current state, a combination of both context vectors is used to guide the decoder to generate more factual words.

Attention mechanism for the facts in the original claim

$$e_{t,i}^{fact} = MLP\left(\mathrm{s_t}, \mathrm{h_i^{fact}}\right) \quad (33)$$

$$\mathrm{a_{t,i}^{fact}} = \frac{\exp\left(\mathrm{e_{t,i}^{fact}}\right)}{\sum_j \exp\left(\mathrm{e_{t,j}^{fact}}\right)} \quad (34)$$

$$\mathrm{c_t^{fact}} = \sum_i \mathrm{a_{t,i}^{fact}}\ \mathrm{h_i^{fact}} \quad (35)$$

Attention mechanism for the generated target-specific based claim

$$\mathrm{e_{t,i}^{gen}} = \mathrm{MLP}\left(\mathrm{s_t}, \mathrm{h_i^{gen}}\right) \quad (36)$$

$$\mathrm{a_{t,i}^{gen}} = \frac{\exp\left(\mathrm{e_{t,i}^{gen}}\right)}{\sum_j \exp\left(\mathrm{e_{t,j}^{gen}}\right)} \quad (37)$$

$$\mathrm{c_t^{gen}} = \sum_i \mathrm{a_{t,i}^{gen}}\ \mathrm{h_i^{gen}} \quad (38)$$

The context vectors are merged by using Eqs. (39) – (42).

$$\mathrm{c_t} = \left[\mathrm{c_t^{fact}}; \mathrm{c_t^{gen}}\right] \quad (39)$$

$$\mathrm{s_t} = \mathrm{GRU}(\mathrm{Y_{t-1}}, \mathrm{c_t}, \mathrm{s_{t-1}}) \quad (40)$$

$$\mathrm{m_i} = \tanh\left(\mathrm{W_1} \cdot \left[[\mathrm{s_t}, \mathrm{c_t}]\right] + \mathrm{b_1}\right) \quad (41)$$

$$\mathrm{output_i} = \mathrm{softmax}(\mathrm{m_i}) = \frac{\exp(\mathrm{m_i})}{\sum_{t=1}^{C} \exp(\mathrm{m_t})} \quad (42)$$

The unfocused hidden state of the decoder is detected in Eqs. (43) – (44).

$$\mathrm{f_t} = \mathrm{c_t^{fact}} - \mathrm{c_t^{gen}}, \quad (43)$$

$$\mathrm{evalt2_{t;}} = \sum_{i=1}^{T^d} \mathrm{f_t}, \mathrm{h_i^d} \quad (44)$$

The scalar training signals from the discriminator is based on Eqs. (45) – (46).

$$\mathrm{Adm1} = \mathrm{sigmoid}\ (\mathrm{W.\ evalt1} + \mathrm{b}) \quad (45)$$

$$\mathrm{Adm2} = \mathrm{sigmoid}\ (\mathrm{W.\ evalt2_2} + \mathrm{b}) \quad (46)$$

Adm1 is the scalar training signal for generator training, which means there is still unfocused generated information; adm1 is fed to target-specific based claim generator helping to reduces the unimportant information and focus more on the main target topic. Adm1 maximises the probability of generating a target-specific based claim toward a target topic. W is the weight matrix, and b is the bias. Adm2 is a scalar training signal for generator training; it ensures that the original article's fact is preserved. Adm2 is fed to the target-specific based claim generator, which helps to avoid changing the fact. Adm2 represents the gap content between the factual and non-factual generated target-specific based claim. The gap guides the generator to preserve the fact. The probability of generating the next word is based on the SoftMax layer result.

3.5. Stance Detection

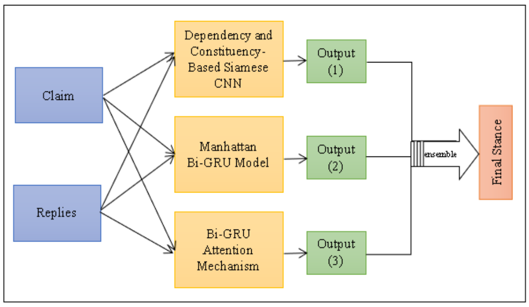

Fig. 5. The proposed Stance Detection Model.

The current model uses three methods to detect the generated target-specific based claim's stance toward all replies, as shown in figure 5. Method 1 uses a dependency and constituency-based Siamese CNN to detect stance, Method 2 uses Manhattan distance to detect stance, and Method 3 uses an attention mechanism to detect stance. To arrive at a final prediction, all possible outcomes of models are considered. The calculation could be performed with each model contributing an equal amount to the ensemble prediction or with each model contributing a different outcome based on its contribution to the ensemble prediction's weighting. The method of weighted ensemble outperforms the method of the equal-weighted ensemble. Ensemble techniques are used to detect stances, which aggregate our baseline classifiers. Ensemble methods are based on the concept of combining multiple models (base classifiers) in order to create a more accurate and reliable model than a single model can provide. As a result, each model is weighted differently.

3.5.1 Method 1 for stance detection: Dependency and constituency-based Siamese CNN

This work extracts sentence-level features from the generated target-specific based claim and replies. It suggests the Stanford Parser be used to perform constituency and dependency parse on the inputs in order to extract the most important sentence information, such as the main linguistic structure that provides information, for example, the subject, predicate, and object of a sentence that are the essential roles in a sentence. It is necessary to learn better sentence representations to observe the structure of sentences and the relationship between the words for each of them.

Word sequences based on their text's constituency words are concatenated with their dependency parser-based word sequences as an input to this work's CNN-based model, where each of them is fed to convolution operations separately for both claim and target-specific based claim. The convolution operation applies a filter w and bias as in Eq. (47) with sigmoid function to words representing constituency or dependency that occur in the sentence, and they are all concatenated to represent the entire feature map for all the words in the sentence. This model combines information for both constituency and dependency-based CNN models to exploit more vital information and capture different features. Concatenate these representations to produce the statement (claim or target-specific based claim) representation, the sum of all vectors. A convolutional neural network is used to transform the generated target-specific based claim representation from word embedding vectors to semantic sentence hidden states. Then, to reduce the representation's spatial size while retaining essential features, a pooling operation is used. The attention vector is generated by connecting the target topic and target-specific based claim representation feature vectors into one vector. The matching distance is used to determine how similar target-specific based claim a and a claim replies [105].

Constituency based CNN for replay X_c

For a given claim, the convolution operation applies a filter W for each constituency based on concatenated word sequence X_c (or words with is constituent structures words from the constituency sub-tree) from the claim where b_c is the bias as follow:

$$\mathrm{Y_{ci}} = \mathrm{sigmoid}(\mathrm{W_c X_c} + \mathrm{b_c}) \qquad (47)$$

All words constituency information is concatenated to generate the feature map const, as in Eq. (48):

$$\mathrm{const} = \mathrm{Y_{c1}}\,, \mathrm{Y_{c2,}} \mathrm{Y_{c3}}\ \mathrm{Y_{cl}} \qquad (48)$$

Dependency-based CNN for replay X_{dep}

For a given claim, the convolution operation applies a filter W for each dependency-based concatenated word sequence X_{dep} (or words with is dependent structures words from the dependency sub-tree) from the claim where b_c is the bias as Eq. (49):

$$\mathrm{Y_{depi}} = \mathrm{sigmoid}(\mathrm{W_c X_{dep}} + \mathrm{b_c}) \qquad (49)$$

All words dependency information is concatenated to generate the feature map dep, Eq. (50):

$$\mathrm{dep} = \mathrm{Y_{dep1}}\,, \mathrm{Y_{dep2}} \mathrm{Y_{dep3}}\ \mathrm{Y_{depl}} \qquad (50)$$

Dependency-based CNN and constituency-based CNN concatenation for replay

The max-pooling operation is applied for both feature maps, const. and dep., to extract the most significant features from each of them, then they are concatenated as Eq. (51):

$$\mathrm{max1} = \max\,(\mathrm{dep}) + \max(\mathrm{const}) \qquad (51)$$

The same equations, Eqs. (47) – (51), used for claim feature representation, will be used for the target-specific based claim, Eqs. (52) – (56).

Constituency-based CNN for the target-specific based claim X_e

$$\mathrm{Y_{ei}} = \mathrm{sigmoid}(\mathrm{W_e X_e} + \mathrm{b_{e)}} \qquad (52)$$

$$\mathrm{e} = \mathrm{Y_{e1}}\,, \mathrm{Y_{e2}} \mathrm{Y_{e3}}\ \mathrm{Y_{en}} \qquad (53)$$

Dependency-based CNN for the target-specific based claim X_{dep2}

$$\mathrm{Y_{depi}} = \mathrm{sigmoid}(\mathrm{W_e X_{depi}} + \mathrm{b_{ce}}) \qquad (54)$$

$$\mathrm{dep2} = \mathrm{Y_{dep1}}\,, \mathrm{Y_{dep2}} \mathrm{Y_{dep3}}\ \mathrm{Y_{depl}} \qquad (55)$$

Dependency-based CNN and constituency-based CNN concatenation for the target-specific based claim

$$\mathrm{max2} = \mathrm{ma\,x}(\mathrm{dep2}) + \max(\mathrm{const}) \qquad (56)$$

After generating the vector representation of sentences, three matching methods are applied to extract relations between ($max1$; $max2$)

1. Concatenation of individual representation ($max1$; $max2$) to produce r1
2. Element-wise product ($max1$* $max2$) to produce r2
3. Absolute element-wise difference ($max1$- $max2$) to produce r3

All the resulting vectors r1, r2, and r3 are concatenated and fed to a SoftMax classifier to predicts the stance label between Claim and target-specific based claim as Eq. (57):

$$\mathrm{f} = \mathrm{r1} \oplus \mathrm{r2} \oplus \mathrm{r3} \qquad (57)$$

f is a fully connected neural network, Eq. (58).

$$\mathrm{output1} = \mathrm{softmax}(\mathrm{W_1 f} + \mathrm{b_1}) \qquad (58)$$

3.5.2. *Method 2 for stance detection: Manhattan- Bi-GRU Model*

Bi-GRU is applied to extract the final hidden state's representation, a vector representation for each claim (replay) and target-specific based claim and then use them to compute the semantic similarity between them. The semantic similarity is computed by the Manhattan Bi-GRU Model [105], as Eq. (59) where the distance is transformed into a similarity score to measure the strength of the claim toward the target-specific based claim. $h^{(c)}$ and $h^{(e)}$ are the last hidden representations for the claim and target-specific based claim respectively.

$$output\ 2 = exp\left(-\|h^{(c)} - h^{(e)}\|_1\right) \quad (59)$$

3.5.3. *Method 3 for stance detection: Bi-GRU Attention mechanism*

Attention mechanisms capture the most relevant features to detect the target-specific based claim's stance toward the primary target topic. This model merges them as one vector after extracting the hidden states for Bi-GRU's target topic and target-specific based claim. The word attention weights are computed using Eqs. (60) – (73), where $h^i_{conc}\ and\ h^i_{claim}$ are the average of hidden states from Bi-GRU for the target-specific based claim and claim, respectively, $\alpha_i\ and\ \beta_i$ are attention vectors for both claim and target-specific based claim that used to compute word attention weights. Then the text representation considers the common features between them as in Eq. (74):

$$\mathrm{att}(\mathrm{h}^{\mathrm{i}}_{\mathrm{conc}}) = \tanh(\mathrm{h}^{\mathrm{i}}_{\mathrm{conc}} \cdot \mathrm{W}_1 + \mathrm{b}_1) \quad (60)$$

$$\alpha_{\mathrm{i}} = \frac{\exp\left(\mathrm{att}(\mathrm{h}^{\mathrm{i}}_{\mathrm{conc}})\right)}{\sum_{\mathrm{j}=1}^{\mathrm{n}+1}\exp\left(\mathrm{att}(\mathrm{h}^{\mathrm{i}}_{\mathrm{conc}})\right)} \quad (61)$$

The following Eq. (62) is applied to generate the final representation for the target-specific based claim representation with Target's topic.

$$\mathrm{target} - \mathrm{specific\ based\ claim}_{\ \mathrm{r}} = \sum_{\mathrm{i}=1}^{\mathrm{n}+1}\alpha_{\mathrm{i}}\,\mathrm{h}^{\mathrm{i}}_{\mathrm{conc}} \quad (62)$$

For the claim concatenated with the target topic, the attention vector is calculated by Eqs. (63) – (66):

$$att\left(h^i_{claim}, h^p_{conc}\right) = tanh\left(h^i_{claim} \cdot h^p_{conc} \cdot W_2 + b_2\right) \quad (63)$$

$$\beta_{\mathrm{i}} = \frac{\exp\left(\mathrm{att}\left(\mathrm{h}^{\mathrm{i}}_{\mathrm{claim}}, \mathrm{h}^{\mathrm{p}}_{\mathrm{conc}}\right)\right)}{\sum_{\mathrm{j}=1}^{\mathrm{m}}\exp\left(\mathrm{att}(\mathrm{h}^{\mathrm{i}}_{\mathrm{claim}}, \mathrm{h}^{\mathrm{p}}_{\mathrm{conc}})\right)} \quad (64)$$

$$\mathrm{claim}_{\mathrm{r}} = \sum_{\mathrm{i}=1}^{\mathrm{m}}\ \beta_{\mathrm{i}}\mathrm{h}^{\mathrm{i}}_{\mathrm{claim}}. \quad (65)$$

$$\mathrm{output3} = \mathrm{softmax}([\mathrm{claim}_{\mathrm{r}} \oplus \mathrm{target} - \mathrm{specific\ based\ claim}_{\ \mathrm{r}}]\ ) \quad (66)$$

The final stance fs of each article is based on the average of output1, output2 and output3. Where w1+w2+w3=1, weights contribute equally 33.33% weight for each of the models Fs=w1* output1+ w2* output2+ w3* output3. The weighted average ensemble is used to improve prediction scores. Fs=w1* output1+ w2* output2+ w3* output3, w1=0.6, w2=0.2, w3=0.2 that have been empirically established and have a higher predictive performance on final prediction

3.6. *Argumentation-based Truth Discovery*

Zhao et al. [22] presented preliminary steps towards truth discovery methods based on bipolar argumentation, where a truth discovery network is mapped to a bipolar argumentation framework by assigning a trust score to each source and a belief score to each claim. Cayrol & Lagasquie-Schiex [106] also suggest linking Truth Discovery with Bipolar Abstract Argumentation. They consider a truth discovery network as disjoint sets S, O and F, representing sources, target topics and claims, respectively. For argumentation frameworks, they consider that arguments interact through attacks and support relations. They provide an example as in figure 6 to illustrate graph representation of a truth discovery network where sources are s, t, u, v, target topics are o, p and claims are f, g, h, i. For the claims related to target topics o and p, the sources s and t have contradiction views toward f, h while the sources u and v agree source s particularly t on target topic p. They propose a truth discovery operator that assigns each source a trust score and each claim a belief score. Argumentation-based Truth Discovery is inspired by abstract argumentation by identifying such arguments with the sources and claims. They propose to encode source trustworthiness by introducing an argument, e.g., "s is a trustworthy source", and introduce an argument "f is a believable claim" for identifying claim believability. According to the example, B(N) yields two meta-arguments where each argument attacks the other: X1, X2, where X1 = {s, f, h} , X2={ t, u, v, g, i}

The proposed work applies an argumentation-based truth discovery, as Fig. 7 and 8, where different arguments support contrary target-specific based claim s for specific information from multiple sources with different degrees of trustworthiness, e.g., f and g are contrary from conflict sources, s and t on a particular target topic p.

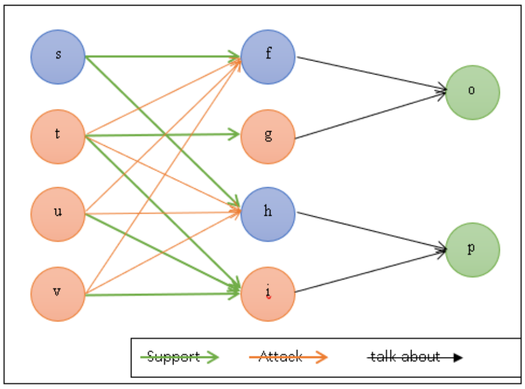

Fig. 6. Graph representation of a truth discovery network [73], s and t disagree on the fact f, h for target topics o and p. Sources u and v do not comment on claim g but agree with t on target topic p.

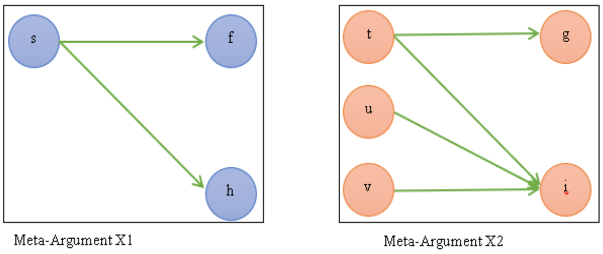

Fig. 7. Argumentation-based truth discovery: meta-arguments X1, X2

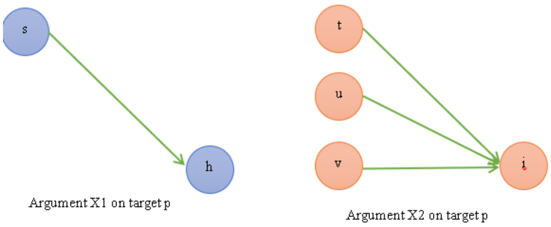

Fig. 8. Argumentation-based truth discovery: arguments X1, X2 on target P

For two meta-arguments X1, X2, where each argument attacks the other, including the sources with their supported claims, estimation source reliability weight is applied to compute the strength of supporting compared to attacking. First, each meta-argument, including sources with its supported claim, is expressed, e.g.,

X1={s,f,h} ,
X2={ t,u,v,g,i}

For each target topic, e.g., on target topic p, the candidate truth claims and reliable sources are put in a set for both attacking arguments, e.g.,

Arg1={s, h}
Arg2={t, u, v, i}

As a result, the strength of the arguments is calculated as follows to select a target topic's truth claim.

1. *Relevance score:* this score measure how the claim (e.g. …) and its supporting replies (e.g. …) are relevant and cover the same issue. For each pair of claim and supporting replies, word embeddings in the same meta-argument, the Siamese adaptation of the Long Short-Term Memory (LSTM): Manhattan LSTM model [105]is used because….This model employs an LSTM, with each claim and source represented by the final hidden state. The semantic similarity between them is then determined. After that, all of this model's outputs are averaged for all claim-source pairs.

2. *The dependency between the data sources scores*: The data sources are interdependent, and there is no conflicting information or inconsistency: conflicting claims frequently prevent people from reaching the same target-specific based claim based on the same evidence. The highest correlated source with other sources is computed in the same way as the claim-source pair. The difference between all sources and its supported sources vector and other vectors from other sources, such as the Manhattan distance, is averaged, compared, and ranked (reliable to unreliable). This model computes the probability of correlating with other sources in each meta-argument; u and v are vectors of different sources, Eq. (67).

   $$p(s_i|v_i,u_i) = \frac{exp(v_i,u_i)}{\sum_i exp(v_i,u_i)} \qquad (67)$$

   If the probability is >=0.5, then the source is selected as a candidate trustworthy source. The sources with more correlation and dependency with other sources are considered as a trustworthy source
3. *Interpretation score:* replay should justify the claim in order to interpret its acceptability. Greater weight is placed on the reasons for accepting a target-specific based claim: the more likely it is that the target-specific based claim is true. This model calculates the probability of each claim in each meta-argument being supported by its sources, Eq. (68):

   $$p(s_i|c_i,u_i) = \frac{exp(c_i,u_i)}{\sum_i exp(c_i,u_i)} \qquad (68)$$

   $p(s_i)$ is the probability of a source u supporting a claim c, i.e., to what extent its associated sources support the claim, u_i is the source vector representation and c_i is the claim vector representation. If the probability is >= 0.5, then the claim is selected as a candidate truth.
4. Sufficient sources: To accept the claim based on a specific target, there must be a sufficient number of relevant and acceptable premises.
5. Conflict sources: That is, this is an argument for controversial issues. A strong argument includes an effective rebuttal to the argument. The provided argument addresses the strongest counterarguments effectively.
6. *Consistency:* Replay embeddings are averaged as a ground truth claim for each meta-argument. The claim with the highest similarity to the average is considered more likely as a truth claim from this argument with its supporting sources.
7. *Argument style:* Finally, as shown in Barrón-Cedeño et al. [107], sufficient vocabulary richness and readability features are used to determine both arguments' most trustworthy source and truth claim. All feature results are weighted to determine the final veracity label. According to Potthast et al. [18], hyperpartisan outlets have a different writing style than mainstream news outlets. Rashkin et al. [108] investigated the relationship between words from the lexicons above in various news articles. They discovered that certain words from their lexicons (swear, see, and negation) appear more frequently in propagandistic, satirical, and hoax articles than in reliable news articles.

Let the index for each argument about a specific target topic be I = 1,..., n. The candidates assert that they have an equal chance of getting it right. For each argument representation, the source and replies are represented by two vectors, content embeddings and processed user profile information embeddings, as Liu et al. (Liu et al., 2015) described. All of the inputs are merged with all of the features described above as input representations. Shared parameters are learned to classify each argument independently, yielding the logits:

Ri= θ [input representations]

Ri … Rn are then concatenated and passed through SoftMax to determine a probability distribution over all arguments.

Given the prediction of all tasks, a global loss function forces the model to minimise the cross-entropy of the prediction and true distributions for all tasks.

$$\mathcal{L} = \sum_{i=1}^{N} \lambda_i L(\hat{y}_i, y_i) \qquad (69)$$

$$L(\hat{y}_i, y_i) = y_i log\hat{y}_i + (1 - y_i)log(1 - \hat{y}_i) \qquad (70)$$

Where λ_i is the weight for task i, and N is the number of tasks, stance detection and veracity prediction.

4. Experiments and Results

The baseline approaches are included to facilitate a thorough comparison of our approach. The experiments aim to answer the following questions:

- RQ1: Can our proposed model ATD outperform the baseline models?

- RQ2: What is the effect of each module in ATD on performance improvement. For example, to what extent could generating the target-specific based claim of an argument for a specific target topic aid in inferring stances from related replies? If you select the informative clauses that correspond to a specific target topic and ignore the noisy clauses, you will come up with a better target-specific based claim that conveys a pro or con stance on replies. This model supposes that it is good to focus on the relevant parts of an article for stance detection and then predict the overall veracity. How important is it for a fact-preserving evaluator and a focused evaluator to produce a better summary of the main target topic? To which extent ensemble learning helps to improve the stance detection results? And finally, to what extent are argumentation theories and comparing the strength of arguments beneficial in predicting the truth?

4.1. Datasets

Different datasets were developed to train the systems for veracity checking, where evidence comes from trusted information sources. Vlachos and Riedel [109] built a dataset with 221 statements and hyperlinks to the evidential source. Other datasets focus more on their information as features, such as metadata on the speaker, in the LIAR dataset [110] with about 12k labelled claims. Recently, the most used dataset is FEVER [42], a large-scale dataset for fact extraction and verification with about 185k labelled claims. For contradictory claims, the SemEval-2019 Task 7 dataset was developed by Gorrell et al. [13]. Finally, the Emergent dataset was developed by Ferreira and Vlachos [14] with 300 claims 2,595 associated news articles.

Macro-averaged F1 has been used as the evaluation metric for the two tasks in RumorEval 2019 [13], as discussed in section 2 above. The RumorEval dataset includes 325 rumorous conversation threads with a training/development/testing split. Additional experiments are conducted on Emergent publicly available datasets [14] as discussed in section 2 above since news article headline and evidence of news article content are essential information for this model training. Since the Emergent dataset is the largest, collected from a fact-checking website, more balanced and annotated to help the model train, this dataset will be used in the experiment. The claims are paired with news headlines and their stances and the public veracity of the claim. Both RumorEval 2019 and Emergent datasets are annotated in some way that aids in the training of our model, but the primary difference is that emergent claims data are longer and have multiple target topics.

In this experiments, Macro-averaged F1 is used to evaluate the performance on both tasks because it solves the imbalanced data problem. The statistical information for Rumour Eval-2019 datasets is described in table 5 and 6. We also evaluate the proposed framework using an additional dataset, Emergent corpus, since headline annotations draw attention to the article where Emergent is a dataset of rumours (claims) coupled with news headlines and their stances. Since our model focuses on generating target-specific based claim s for news articles and summarising a long article into a target-specific based claim, it uses extra information, like the headline in Emergent data that represents the news store, to increase the accuracy. The statistical information for the Emergent datasets is illustrated in tables 7 and 8.

Table 5

Rumour Eval-2019 Task A stance detection corpus [13].

	Support	**Deny**	**Query**	**Comment**	**Total**
Twitter Train Reddit Train	1004 23	415 45	464 51	3685 1015	5568 1134
Total Train	1027	460	515	4700	6702
Twitter Test Reddit Test	141 16	92 54	62 31	771 705	1066 806
Total Test	157	146	93	1476	1872
Total Task A	1184	606	608	6176	8574

Table 6

Rumour Eval-2019 Task B veracity checking corpus [13].

	True	False	Unverified	Total
Twitter Train Reddit Train	145 9	74 24	106 7	325 40
Total Train	154	98	113	365
Twitter Test Reddit Test	22 9	30 10	4 6	56 25
Total Test	31	40	10	81
Total Task B	185	138	123	446

Table 7

Emergent dataset [14].

Claims	300
Headlines	2,595
Minimum number of articles per claim	1
Maximum number of articles per claim	50
Training instances	2,071
Test instances	524

Table 8

Statistics of the Emergent dataset.

Subject	Stance	Emergent Number	Percentage
Training	agree	992	24.37
	disagree	303	7.44
	discuss	776	19.06
	unrelated	2,000	49.13
		4,071	
Testing	Agree	246	24.02
	disagree	91	8.89
	discuss	776	19.06
	unrelated	500	48.83
		1,024	

4.2. Experimental Setting

In the experiments, the models in this paper are implemented using Keras. All word vectors are initialised using word2vec [100] and Glove [27], Where we discover that Glove performs better than word2vec. The hyperparameters, variables set before training which values are used to control the learning process and before optimizing the weights and bias, are chosen to achieve the most considerable value on the validation set and then train the model on the entire dataset. In our implementation, the word embedding dimension is 300, the size of hidden units in GRU is 100. The batch size is set 32. The learning rate is set at 0.001. The rule activation function used in the hidden layers is set to evaluate Task A and B's performance. The used evaluation metric: "macro-average F1 score" since the class labels are imbalanced.

4.3. Baseline Comparison

In this section, the performance of stance detection and rumour verification for the proposed model against the state-of-the-art model is discussed in this section.

4.3.1 Emergent dataset training and testing

Our model is compared against the state-of-the-art model reported in Zhang et al. [25] on the augmented Emergent dataset. In addition, experiments are performed on the publicly available Emergent dataset [14], consisting of news article headlines, evidence of news article content and stances. Stances can be classified as support, oppose, and discuss, which can infer veracity.

The results are shown in Table 9. For veracity detection, our model obtains an accuracy of 78.83%. Since most previous models on the Emergent dataset focus only on the stance detection task but the veracity task, no comparison with the baseline is made.

Table 9

Performance comparison of the model against the State-of-the-Art model [25] for stance detection task on Emergent dataset

Model	Accuracy (%)		
	agree	disagree	discuss
Zhang et al's model [25]	82.52	69.05	84.30
Our model	**83.12**	**73.89**	**89.13**

4.3.2 RumorEval 2019 dataset training and testing

As discussed in section 2 above, Li et al.'s [7] and Khandelwal's [39] models show better performance compared with the top-5 systems in RumourEval 2019 [35]. Li et al.'s model [7], achieves the best performance for veracity detection, but they did not present results for stance detection as a single task. Regarding stance detection results, the best results are shown in Khandelwal [39], so that to evaluate this work's stance detection, a comparison is made with Khandelwal [39] model as shown in table 10 and, for veracity checking, the comparison with Li et al. [7] as shown in table 11.

Table 10

Test results for Task A: stance detection on RumorEval

The model	Macro-F
Khandelwal's[39] Method – Top N_s using (A + B + C)	0.672
Our proposed model	**0.695**

Table 11

Test results for Task B: rumour veracity on RumorEval

The model	**Macro-F**
Li et al.’s model[7]	0.606
Our proposed model	**0.647**

For task A, stance detection, the work in Khandelwal [39] achieves the best Macro-f of 0.6720 and our model achieves 0.695, while for task B, veracity checking, the work in Li et al. [7] achieves the best Macro-f of 0.606 and our model achieves 0.647.

4.4. Ablation study

By examining research question RQ2, ablation tests are conducted on the target topic extraction module, relevant clause retrieval and target-specific based claim generation, weighted ensemble learning for stance detection, and argumentation-based truth discovery; the results are shown in tables 12 and 13. As a result, each module significantly improves overall performance, demonstrating its efficacy.

4.5. Discussions

From the results given above, it is clear that the proposed method shows the best performance among these models. The proposed model outperforms both tasks, achieving Macro F1 0.695 for Task A and 0.647 for task B of the RumorEval 2019 dataset. The remainder of the subsection provides an analysis and evaluation of the proposed model and the results.

First, training this model with and without applying the first component: the primary claim target topic extraction. The performance results revealed that this guide focuses more on the claim's primary target topic and positively contributes to stance classification performance. Target topic-clause retrieving help to ignore noise information as the noise may give wrong indications to deceive the model. This model is trained to classify stances without considering the target topic information, and a decreased accuracy is obtained. To show that the stance and rumour detection benefit from target topic aware target-specific based claim, experiments are conducted to detect evidence against claims without making a target-specific based claim based on the target topic. The change of macro-F1 scores on the two datasets shows the improvements by capturing certain words related to the target topic and eliminating the irrelevant. The macro-F1 score is chosen as a metric to give each task equal weight because it resolves the data imbalance. It outperforms the previous best baseline methods for the Emergent data. This could be that the model detects the article's stance against a claim by paying more attention to the claimed target topic, while the original article may have various target topics to talk about it. It is observed that word alignment can capture the target topic information for better performance of stance detection as target topic-specific attention provides more concise information, discarding another target topic the claim does not concern with it.

The results emerging from these experiments confirmed the effectiveness of generating target-specific based claim conditioned on the target topic representation that is finally presented to the target topic claim and showed that it could be useful by extracting salient information from a long article without including less salient information., A significant improvement in this model's general results on both tasks A and B is achieved. Compared with baselines for stance detection, the advantage of knowledgeable target topic and target-specific based claim is demonstrated. A significant improvement on an emergent dataset from 82.52 %, 69.05 %, 84.30 % in Zhang et al’s model [25] , to 83.12%, 73.89%, 89.13% for the three stance labels respectively as illustrated in table 9.

To investigate the applicability of the proposed model on new unseen data, where there is no knowledge related to this event, truth discovery is very beneficial to generalise veracity prediction since it depends on estimation without supervision. Despite unobserved samples, they may have semantic and syntactic features to that unseen news. The proposed model works well for the different text from two different datasets.

The following observations have been made based on the ablation experiments, as in table 12 and table 13:

- For the Emergent data, the veracity detection accuracy decreases when the target-specific based claim generation is not considered. This is particularly the case for a long article since the headline captures the primary information in making first impressions to readers.
- When the generated target-specific based claim does not cover the target topic of the claim or the extracted target topic is not valid, the performance is decreased
- Models augmented with truth discovery perform better than those without, i.e., assigning more scores to the claims inferred by more trusty sources.

- A significant improvement in integrating both tasks stance detection with rumour prediction.
- Since the sources' trustworthiness is not available and there is no prior information, this work's method can significantly enhance reliability source inference by estimating the trust based on Argumentation-based Truth Discovery.
- Utilising the claimed target topic helps the generator produce a concise target-specific based claim, and the evaluator can narrow the cosine distance.
- unlike that most of the previous studies as discussed earlier that either detect stance detection without considering the target topic or focused on inferring the stance for a set of predefined target topics [111], our model extracts a specific target topic to predict the stance toward it separately from stance classification,
- For size limitation, deep learning models need a high volume of data for training; it requires larger datasets than currently available, so this model is expected to perform better if more samples are obtained.
- Sometimes the e model fails to predict some stance labels correctly, maybe due to the lack of current information and other external evidence, e.g., the warrant is needed, so merging them may make additional enhancements, especially in the case.

5. Conclusion

A multi-task learning framework for jointly predicting rumour stance and veracity is proposed, where the source reliability is considered. A new deep learning model with a novel architecture is designed and studied to discover multiple truths from conflicting sources by connecting truth discovery methods with bipolar argumentation. The experiments with two influential datasets show that the proposed model outperforms state-of-the-art stance detection and rumour verification tasks. Argumentation-based truth discovery provides an effective way towards veracity detection by discovering the acceptable arguments through reframing truth discovery in terms of argumentation; this implies describing the arguments and the attack and support relations. There are several ways to move the current work forward. The current work involves source-claim and source-source relationships and focuses on information richness in order to obtain confident score information. We are also planning to modify this model by considering other argumentation components such as warrant and backing in the Toulmin model and consider other factors like the source reputations.

Table 12

Ablation experiment of our model, stance detection scores of different ablation models.

Model	On emergent (relative score)	On RumourEval Macro-f
ATD without target topic extraction	82.19	0.623
ATD without target-specific based claim generation	79.36	0.679
ATD without clause relevant retrieving for target-specific based claim generation	83.92	0.678
ATD without evaluators for target-specific based claim generation	82.64	0.685
ATD without weighted ensemble learning for stance detection (i.e., each model contributes equally)	84.72	0.627
ATD without argumentation-based truth discovery	85.84	0.635
ATD with argumentation-based truth discovery	**89.97**	**0.695**

Table 13

Rumour veracity of scores of different ablation models.

Model	On emergent (relative score)	On RumourEval Macro-f
ATD without target topic extraction	75.77	0.603
ATD without target-specific based claim generation	72.08	0.639
ATD without clause relevant retrieving for target-specific based claim generation	73.83	0.636
ATD without evaluators for target-specific based claim generation	74.83	0.631
ATD without weighted ensemble learning for stance detection (i.e., each model contributes equally)	72.83	0.576
ATD without argumentation-based truth discovery	71.83	0.583
ATD with argumentation-based truth discovery	**78.83**	**0.647**

References

[1] S. A. Alkhodair, S. H. H. Ding, B. C. M. Fung, and J. Liu, “Detecting Breaking News Rumors of EmergingTopics in Social Media,” *Inf. Process. Manag.*, vol. 57, no. 2, p. 102018, 2020, doi: 10.1016/j.ipm.2019.02.016.

[2] J. Harsin, “[Proto-Post-truth] The Rumour Bomb: Theorizing the Convergence of New and Old Trends in Mediated US Politics,” *South. Rev. Commun. Polit. Cult.*, vol. 39, no. 1, 2006, pp. 84–110, 2018, [Online]. Available: https://search.informit.com.au/documentSummary;dn=264848460677220;res=IELAPA.

[3] S. Vosoughi, D. Roy, and S. Aral, “The Spread of True and False News Online,” *Science (80-.).*, vol. 1151, no. March, pp. 1146–1151, 2018, doi: 10.1126/science.aap9559.

[4] K. Shu, H. R. Bernard, and H. Liu, “Studying Fake News via Network Analysis: Detection and Mitigation,” in *Nitin Agarwal, Nima Dokoohaki, Serpil Tokdemir: Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, Springer, Cham, 2019, pp. 43–65.

[5] Z. Jin, J. Cao, Y. Zhang, and J. Luo, “News Verification by Exploiting Conflicting Social Viewpoints in Microblogs,” in *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 2016, pp. 2972–2978.

[6] Q. Li, Q. Zhang, L. Si, and Y. Liu, “Rumor detection on social media: Datasets, methods and opportunities,” in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, 2019, pp. 66–75, doi: 10.18653/v1/d19-5008.

[7] Q. Li, Q. Zhang, and L. Si, “Rumor Detection By Exploiting User Credibility Information, Attention and Multi-task Learning,” in *Proceedings ofthe 57th Annual Meeting ofthe Association for Computational Linguistics*, 2020, pp. 1173–1179, doi: 10.18653/v1/p19-1113.

[8] Q. Li, X. Liu, R. Fang, A. Nourbakhsh, and S. Shah, *User Behaviors in Newsworthy Rumors : A Case Study of Twitter*, no. ICWSM. Association for the Advancement of Artificial Intelligence (www.aaai.org), 2016, pp. 627–630.

[9] Y. Liu and Y. F. B. Wu, “Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32, no. 1, pp. 354–361, [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11268.

[10] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, “Detection and Resolution of Rumours in Social Media: A Survey,” *ACM Comput. Surv.*, vol. 51, no. 2, pp. 1–36, 2018, doi: 10.1145/3161603.

[11] M. Mohtarami, R. Baly, J. Glass, P. Nakov, L. Màrquez, and A. Moschitti, “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)*, 2018, vol. 1, pp. 767–776, doi: 10.18653/v1/N18-1070.

[12] J. Ma, W. Gao, and K. Wong, “Rumor Detection on Twitter with Tree-structured Recursive Neural Networks,” in *Proceedings ofthe 56th Annual Meeting ofthe Association for Computational Linguistics (Long Papers)*, 2018, pp. 1980–1989, doi: 10.18653/v1/P18-1184.

[13] G. Gorrell, K. Bontcheva, L. Derczynski, E. Kochkina, M. Liakata, and A. Zubiaga, “RumourEval 2019: Determining Rumour Veracity and Support for Rumours,” in *Proceedings ofthe 13th International Workshop on Semantic Evaluation (SemEval-2019)*, 2019, pp. 845–854.

[14] W. Ferreira and A. Vlachos, “Emergent: A Novel Data-set for Stance Classification,” in *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, 2016, vol. June 12-17, no. 1, pp. 1163–1168, doi: 10.18653/v1/n16-1138.

[15] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media: A Data Mining Perspective,” *ACM SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, 2017, doi: 10.1145/3137597.3137600.

[16] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini, “Computational fact checking from knowledge networks,” *PLoS One*, vol. 10, no. 6, pp. 1–13, 2015, doi: 10.1371/journal.pone.0128193.

[17] N. J. Conroy, V. L. Rubin, and Y. Chen, “Automatic deception detection: Methods for finding fake news,” in *Proceedings of the Association for Information Science and Technology-ASIST*, 2015, vol. 52, no. 1, pp. 1–4, doi: 10.1002/pra2.2015.145052010082.

[18] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, “A Stylometric Inquiry into Hyperpartisan and Fake News,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 2018, vol. 1, no. July, pp. 231–240, doi: 10.18653/v1/p18-1022.

[19] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, “Prominent Features of Rumor Propagation in Online Social Media,” in *Proceedings - IEEE 13th International Conference on Data Mining, ICDM*, 2013, pp. 1103–1108, doi: 10.1109/ICDM.2013.61.

[20] B. D. Horne and S. Adali, “This Just In-Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News,” in *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, 2017, vol. 11, no. 1, pp. 40–49, doi: 10.18653/v1/w18-5507.

[21] F. Yang, A. Mukherjee, and E. Gragut, “Satirical News Detection and Analysis using Attention Mechanism and Linguistic Features,” in *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2017, pp. 1979–1989, doi: 10.18653/v1/d17-1211.

[22] Z. Zhao, P. Resnick, and Q. Mei, “Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts,” in *Proceedings of the 24th i International World Wide Web Conference*, 2015, pp. 1395–1405, doi: 10.1145/2736277.2741637.

[23] A. Zubiaga, G. Wong Sak Hoi, M. Liakata, and R. Procter, “PHEME Dataset of Rumours and Non-rumours,” *figshare. Dataset*, 2016. https://figshare.com/articles/dataset/PHEME_dataset_of_rumours_and_non-rumours/4010619/1 (accessed Nov. 20, 2020).

[24] T. Saikh, A. Anand, A. Ekbal, and P. Bhattacharyya, “A Novel Approach Towards Fake News Detection: Deep Learning Augmented with Textual Entailment Features,” in *International Conference on Applications of*

Natural Language to Information Systems, 2019, vol. 11608 LNCS, pp. 345–358, doi: 10.1007/978-3-030-23281-8_30.

[25] Q. Zhang, S. Liang, A. Lipani, Z. Ren, and E. Yilmaz, "From Stances' Imbalance to Their Hierarchical Representation and Detection," in *Proceedings of the World Wide Web Conference, WWW 2019*, 2019, vol. May 13-17, no. 1, pp. 2323–2332, doi: 10.1145/3308558.3313724.

[26] I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva, "Stance Detection with Bidirectional Conditional Encoding," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 876–885, doi: 10.18653/v1/d16-1084.

[27] Jeffrey Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, no. October, pp. 1532–1543, doi: 10.3115/v1/D14-1162.

[28] S. Bajaj, "' The Pope Has a New Baby !' Fake News Detection Using Deep Learning," *CS 224N*, pp. 1–8, 2017, [Online]. Available: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2710385.pdf.

[29] Q. Zhang, E. Yilmaz, and S. Liang, "Ranking-based Method for News Stance Detection," in *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, 2018, pp. 41–42, doi: 10.1145/3184558.3186919.

[30] A. Hanselowski, A. Pvs, B. Schiller, and F. Caspelherr, "Description of the System Developed by Team Athene in the FNC-1," *Technical report*, 2017. https://github.com/hanselowski/athene_system/blob/master/system_description_athene.pdf (accessed Nov. 24, 2020).

[31] P. Bourgonje, J. Moreno Schneider, and G. Rehm, "From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles," in *Proceedings ofthe 2017 EMNLP Workshop on Natural Language Processing meets Journalism*, 2017, pp. 84–89, doi: 10.18653/v1/w17-4215.

[32] B. Riedel, I. Augenstein, G. P. Spithourakis, and S. Riedel, "A simple but tough-to-beat baseline for the Fake News Challenge stance detection task," *arXiv Prepr. arXiv1707.03264*, pp. 1–6, 2018, [Online]. Available: https://arxiv.org/pdf/1707.03264.pdf.

[33] X. Wang, C. Yu, S. Baumgartner, and F. Korn, "Relevant Document Discovery for Fact-Checking Articles," in *Companion Proceedings of the The Web Conference 2018, International World Wide Web Conferences Steering Committee*, 2018, pp. 525–533, doi: 10.1145/3184558.3188723.

[34] R. Yang, W. Xie, C. Liu, and D. Yu, "BLCU_NLP at SemEval-2019 Task 7: An Inference Chain-based GPT Model for Rumour Evaluation," in *Proceedings ofthe 13th International Workshop on Semantic Evaluation (SemEval-2019)*, 2019, no. June 6-7, pp. 1090–1096, doi: 10.18653/v1/s19-2191.

[35] M. Fajcik, L. Burget, and P. Smrz, "BUT-FIT at SemEval-2019 Task 7: Determining the Rumour Stance with Pre-Trained Deep Bidirectional Transformers," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 1097–1104, doi: 10.18653/v1/s19-2192.

[36] I. Baris, L. Schmelzeisen, and S. Staab, "CLEARumor at SemEval-2019 Task 7: ConvoLving ELMo against rumors," in *3th International Workshop on Semantic Evaluation*, 2019, no. June 06-07, doi: 10.18653/v1/s19-2193.

[37] A. Radford, K. Narasimhan, A. Tim Salimans, and I. Sutskever, "Improving Language Understanding with Unsupervised Learning," *Technical report, Open AI*, 2018. https://openai.com/blog/language-unsupervised/ (accessed Nov. 20, 2020).

[38] M. Peters *et al.*, "Deep Contextualized Word Representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Long Papers)*, 2018, vol. 1, pp. 2227–2237, doi: 10.18653/v1/N18-1202.

[39] A. Khandelwal, "Fine-Tune Longformer for Jointly Predicting Rumor Stance and Veracity," *arXiv Prepr.*, 2020, doi: 10.1145/3430984.3431007.

[40] Q. Li, Q. Zhang, and L. Si, "eventAI at SemEval-2019 Task 7: Rumor Detection on Social Media by Exploiting Content, User Credibility and Propagation Information," in *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, 2019, pp. 855–859, doi: 10.18653/v1/s19-2148.

[41] P. Wei, N. Xu, and W. Mao, "Modeling Conversation Structure and Temporal Dynamics for Jointly Predicting Rumor Stance and Veracity," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2019, pp. 4787–4798, doi: 10.18653/v1/d19-1485.

[42] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a Large-scale Dataset for Fact Extraction and VERification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Long Papers)*, 2018, vol. 1, no. June, pp. 809–819, doi: 10.17863/CAM.40620.

[43] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional Attention Flow for Machine Comprehension," in *ICLR 2017 conference submission*, 2016, pp. 1–13, [Online]. Available: http://arxiv.org/abs/1611.01603.

[44] Y. Nie, H. Chen, and M. Bansal, "Combining Fact Extraction and Verification with Neural Semantic Matching Networks," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 2019, vol. 33, no. 01, pp. 6859–6866, doi: 10.1609/aaai.v33i01.33016859.

[45] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *NAACL -HLT 2019 - 2019 Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, vol. 1, no. Mlm, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[46] A. Soleimani, C. Monz, and M. Worring, "BERT for Evidence Retrieval and Claim Verification," *J. M. Jose al. ECIR 2020, LNCS 12036*, no. 3, pp. 359–366, 2020, doi: 10.1007/978-3-030-45442-5.

[47] O. Enayet and S. R. El-Beltagy, "NileTMRG at SemEval-2017 Task 8: Determining Rumour and Veracity Support for Rumours on Twitter.," in *Proceedings ofthe 11th International Workshop on Semantic Evaluations (SemEval-2017)*, 2017, pp. 470–474, doi: 10.18653/v1/s17-2082.

[48] B. Yorgancı, "Multi-sourced Information

Trustworthiness Analysis: Applications and Theory," State University of New York at Buffalo, 2018.

[49] J. Li, X. Hu, J. Tang, and H. Liu, "Unsupervised Streaming Feature Selection in Social Media," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015, no. October 13, pp. 1041–1050, doi: http://dx.doi.org/10.1145/2806416.2806501.

[50] M. Glenski, T. Weninger, and S. Volkova, "Identifying and Understanding User Reactions to Deceptive and Trusted Social News Sources," *ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 2, no. 1, pp. 176–181, 2018, doi: 10.18653/v1/p18-2029.

[51] M. Mendoza, B. Poblete, and C. Castillo, "Twitter Under Crisis: Can we trust what we RT?," in *SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics*, 2010, pp. 71–79, doi: 10.1145/1964858.1964869.

[52] R. Procter, A. Voss, and F. Vis, "Reading the Riots on Twitter: Methodological Innovation for the Analysis of Big Data," *Int. J. Soc. Res. Methodol. Comput. Soc. Sci. Res. Strateg. Des. Methods*, vol. 16, no. 3, pp. 197–214, 2013, doi: 10.1080/10439463.2013.780223.

[53] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it Identifying Misinformation in Microblogs," in *Proceedings ofthe 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, no. July 27-31, pp. 1589–1599, [Online]. Available: https://www.aclweb.org/anthology/D11-1147.pdf.

[54] Q. Zhang, S. Zhang, J. Dong, J. Xiong, and X. Cheng, "Automatic Detection of Rumor on Social Network," *NLPCC2015, Nat. Lang. Process. Chinese Comput. Springer, Cham*, vol. LNAI 9362, pp. 113–122, 2015, doi: 10.1007/978-3-319-25207-0_10.

[55] A. Zubiaga, E. Kochkina, M. Liakata, R. Procter, and M. Lukasik, "Stance Classification in Rumours as a Sequential Task Exploiting the Tree Structure of Social Media Conversations," in *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, 2016, pp. 2438–2448, [Online]. Available: https://www.aclweb.org/anthology/C16-1230.

[56] M. E. Peters and A. Cohan, "Longformer: The Long-Document Transformer," *arXiv Prepr. arXiv2004.05150v1*, 2020.

[57] E. Kochkina, M. Liakata, and A. Zubiaga, "All-in-one: Multi-task Learning for Rumour Verification," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 3402–3413, [Online]. Available: https://www.aclweb.org/anthology/C18-1288.

[58] J. Ma, W. Gao, and K. Wong, "Detect Rumor and Stance Jointly by Neural Multi-task Learning," in *Proceedings of the Web Conference (WWW 2018 Companion)*, 2018, pp. 585–593, doi: 10.1145/3184558.3188729.

[59] L. Poddar, W. Hsu, M. L. Lee, and S. Subramaniyam, "Predicting Stances in Twitter Conversations for Detecting Veracity of Rumors: a Neural Approach," in *Proceedings of the 30th International Conference on Tools with Artificial Intelligence, ICTAI*, 2018, vol. 2018-Novem, pp. 65–72, doi: 10.1109/ICTAI.2018.00021.

[60] Y. Li *et al.*, "A Survey on Truth Discovery," *ACM SIGKDD Explor. Newsl.*, vol. 17, no. 2, pp. 1–16, 2016, doi: 10.1145/2897350.2897352.

[61] J. Singleton, "Truth Discovery : Who to Trust and What to Believe," in *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020)*, 2020, no. May 9-13, pp. 2211–2213.

[62] J. Singleton, "On the Link Between Truth Discovery and Bipolar Abstract Argumentation," *Online Handb. Argumentation AI*, vol. 1, no. June, pp. 43–47, 2020.

[63] J. Pasternack and D. Roth, "Knowing What to Believe (when you already know something)," in *Proceedings ofthe 23rd International Conference on Computational Linguistics (Coling 2010)*, 2010, vol. 2, no. August, pp. 877–885.

[64] Y. Li *et al.*, "On the Discovery of Evolving Truth," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, vol. 2015-Augus, pp. 675–684, doi: 10.1145/2783258.2783277.

[65] F. Ma *et al.*, "FaitCrowd: Fine Grained Truth Discovery for Crowdsourced Data Aggregation," in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, vol. 2015-Augus, pp. 745–754, doi: 10.1145/2783258.2783314.

[66] J. Marshall, A. Argueta, and D. Wang, "A Neural Network Approach for Truth Discovery in Social Sensing," in *Proceedings - 14th IEEE International Conference on Mobile Ad Hoc and Sensor Systems, MASS 2017*, 2017, pp. 343–347, doi: 10.1109/MASS.2017.26.

[67] X. Yin, J. Han, and P. S. Yu, "Truth Discovery with Multiple Conflicting Information Providers on the Web," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 6, pp. 796–808, 2008, doi: 10.1109/TKDE.2007.190745.

[68] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Integrating Conflicting Data: The Role of Source Dependence," in *Proceedings of the VLDB Endowment*, 2009, vol. 2, no. 1, pp. 550–561, doi: 10.14778/1687627.1687690.

[69] J. Pasternack and D. Roth, "Making Better Informed Trust Decisions with Generalized Fact-Finding," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence Making*, 2011, pp. 2324–2329, doi: 10.5591/978-1-57735-516-8/IJCAI11-387.

[70] A. Galland, S. Abiteboul, A. Marian, and P. Senellart, "Corroborating Information from Disagreeing Views," in *Proceedings of the third ACM International Conference on Web Search and Data Mining (WSDM)*, 2010, pp. 131–140, doi: 10.1145/1718487.1718504.

[71] F. Li, M. L. Lee, and W. Hsu, "Entity Profiling with Varying Source Reliabilities," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, no. August, pp. 1146–1155, doi: 10.1145/2623330.2623685.

[72] C. Dai, D. Lin, E. Bertino, and M. Kantarcioglu, "An Approach to Evaluate Data Trustworthiness Based on Data Provenance," in *Proceedings of the 5th VLDB Workshop on Secure Data Management*, 2008, pp. 82–98, doi: 10.1007/978-3-540-85259-9_6.

[73] X. Dong *et al.*, "Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 601–610, doi: 10.1145/2623330.2623623.

[74] Q. Li *et al.*, "A Confidence-Aware Approach for Truth Discovery on Long-Tail Data," *Proc. VLDB*

Endow., vol. 8, no. 4, pp. 425–436, 2014, doi: 10.14778/2735496.2735505.

[75] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han, "Resolving Conflicts in Heterogeneous Data by Truth Discovery and Source Reliability Estimation," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (SIGMOD '14)*, 2014, pp. 1187–1198, doi: 10.1145/2588555.2610509.

[76] D. Zhou, J. C. Platt, S. Basu, and Y. Mao, "Learning from the Wisdom of Crowds by Minimax Entropy," in *Proceedings of the 25th International Conference on Neural Information Processing Systems NIPS*, 2012, vol. 2, no. December, pp. 2195–2203, [Online]. Available: http://dblp.uni-trier.de/db/conf/nips/nips2012.html#ZhouPBM12.

[77] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On Truth Discovery in Social Sensing: A Maximum Likelihood Estimation Approach," in *IPSN'12 - Proceedings of the 11th International Conference on Information Processing in Sensor Networks*, 2012, no. April, pp. 233–244, doi: 10.1145/2185677.2185737.

[78] M. Samadi, P. Talukdar, M. Veloso, and M. Blum, "ClaimEval: Integrated and Flexible Framework for Claim Evaluation Using Credibility of Sources," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016, no. February, pp. 222–228, [Online]. Available: http://dblp.uni-trier.de/db/conf/aaai/aaai2016.html#SamadiTVB16.

[79] N. Nakashole and T. M. Mitchell, "Language-Aware Truth Assessment of Fact Candidates," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014*, 2014, vol. 1, pp. 1009–1019, doi: 10.3115/v1/p14-1095.

[80] X. Wang, Q. Z. Sheng, X. S. Fang, X. Li, X. Xu, and L. Yao, "Approximate Truth Discovery via Problem Scale Reduction," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15*, 2015, vol. 19-23-Oct-, pp. 503–512, doi: 10.1145/2806416.2806444.

[81] S. Zhi *et al.*, "Modeling Truth Existence in Truth Discovery," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, vol. 2015-Augus, pp. 1543–1552, doi: 10.1145/2783258.2783339.

[82] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han, "A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration," in *Proceedings of the VLDB Endowment (PVLDB)*, 2012, vol. 5, no. 6, pp. 550–561, doi: 10.14778/2168651.2168656.

[83] B. Zhao and J. Han, "A Probabilistic Model for Estimating Real-valued Truth from Conflicting Sources," 2012.

[84] L. Li, B. Qin, W. Ren, and T. Liu, "Truth Discovery with Memory Network," *Tsinghua Sci. Technol.*, vol. 22, no. 6, pp. 609–618, 2017, doi: 10.23919/TST.2017.8195344.

[85] K. Broelemann, T. Gottron, and G. Kasneci, "Restricted Boltzmann Machines for Robust and Fast Latent Truth Discovery," *CoRR*, vol. abs/1801.0, 2018, [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1801.html#abs-1801-00283.

[86] N. Choudhary, R. Singh, I. Bindlish, and M. Shrivastava, "Neural Network Architecture for Credibility," in *In the proceedings of CICLING 2018*, 2018, pp. 1–13, [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1803.html#abs-1803-10547.

[87] L. Xiao, H. Zhang, and W. Chen, "Gated Multi-Task Network for Text Classification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, vol. 2, pp. 726–731, doi: 10.18653/v1/n18-2114.

[88] B. Keller, A. Labrique, K. M. Jain, A. Pekosz, and O. Levine, "Incorporating Copying Mechanism in Sequence-to-Sequence Learning," *J. Med. Internet Res.*, vol. 16, no. 1, p. e8, 2014, [Online]. Available: https://www.jmir.org/2014/1/e8/.

[89] A. See, P. J. Liu, and C. D. Manning, "Get To The Point: Summarization with Pointer-Generator Networks," in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2017, vol. 1, pp. 1073–1083, doi: 10.18653/v1/P17-1099.

[90] W. T. Hsu, C. K. Lin, M. Y. Lee, K. Min, J. Tang, and M. Sun, "A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss," in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2018, vol. 1, pp. 132–141, doi: 10.18653/v1/p18-1013.

[91] Q. Zhou, N. Yang, F. Wei, and M. Zhou, "Selective Encoding for Abstractive Sentence Summarization," in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2017, vol. 1, pp. 1095–1104, doi: 10.18653/v1/P17-1101.

[92] Y. C. Chen and M. Bansal, "Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting," in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2018, vol. 1, no. 2017, pp. 675–686, doi: 10.18653/v1/p18-1063.

[93] X. Chen, S. Gao, C. Tao, Y. Song, D. Zhao, and R. Yan, "Iterative Document Representation Learning Towards Summarization with Polishing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 2020, pp. 4088–4097, doi: 10.18653/v1/d18-1442.

[94] K. M. Hermann *et al.*, "Teaching Machines to Read and Comprehend," in *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15)*, 2015, pp. 1693–1701, [Online]. Available: http://dblp.uni-trier.de/db/conf/nips/nips2015.html#HermannKGEKSB15.

[95] G. Heinrich, "Parameter Estimation for Text Analysis," 2009. [Online]. Available: http://www.arbylon.net/publications/text-est.pdf.

[96] T. M. COVER, *Elements of Information Theory*. John Wiley & Sons, 1999.

[97] A. Mehri, M. Jamaati, and H. Mehri, "Word Ranking in a Single Document by Jensen-Shannon Divergence," *Phys. Lett. Sect. A Gen. At. Solid State Phys.*, vol. 379, no. 28–29, pp. 1627–1632, 2015, doi: 10.1016/j.physleta.2015.04.030.

[98] R. J. Gallagher, A. J. Reagan, C. M. Danforth, and P. S. Dodds, "Divergent discourse between protests and counter-protests: #BlackLivesMatter and #AllLivesMatter," *PLoS One*, vol. 13, no. 4,

pp. 1–23, 2018, doi: 10.1371/journal.pone.0195644.

[99] S. Gao *et al.*, “Abstractive Text Summarization by Incorporating Reader Comments,” *33rd AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 6399–6406, 2019, doi: 10.1609/aaai.v33i01.33016399.

[100] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” in *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2013, vol. 2, pp. 3111–3119, [Online]. Available: https://dl.acm.org/doi/10.5555/2999792.2999959 .

[101] D. Bahdanau, K. H. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015, pp. 1–15.

[102] P. Nema, M. M. Khapra, A. Laha, and B. Ravindran, “Diversity driven Attention Model for Query-based Abstractive Summarization,” in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2017, vol. 1, pp. 1063–1072, doi: 10.18653/v1/P17-1098.

[103] J. Hasselqvist, N. Helmertz, and M. Kågebäck, “Query-Based Abstractive Summarization Using Neural Networks,” *CoRR, abs/1712.06100*, 2017, [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1712.html#abs-1712-06100.

[104] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” in *Proceedings of the ACL-04 Workshop on Text Summarization Branches Out (WAS 2004)*, 2004, no. July, pp. 74–81, [Online]. Available: https://www.aclweb.org/anthology/W04-1013.

[105] J. Mueller and A. Thyagarajan, “Siamese Recurrent Architectures for Learning Sentence Similarity,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2016, no. November, pp. 2786–2792, [Online]. Available: http://dblp.uni-trier.de/db/conf/aaai/aaai2016.html#MuellerT16.

[106] C. Cayrol and M. C. Lagasquie-Schiex, *Gradual Valuation for Bipolar Argumentation Frameworks*, vol. 3571 LNAI, no. June. Berlin, Heidelberg: Springer, Berlin, Heidelberg, 2005.

[107] A. Barrón-Cedeño, I. Jaradat, G. Da San Martino, and P. Nakov, “Proppy: Organizing the News Based on Their Propagandistic Content,” *Inf. Process. Manag.*, vol. 56, no. 5, pp. 1849–1864, 2019, doi: 10.1016/j.ipm.2019.03.005.

[108] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, “Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking,” in *Proceedings ofthe 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2931–2937, doi: 10.18653/v1/d17-1317.

[109] A. Vlachos and S. Riedel, “Fact Checking: Task definition and dataset construction,” in *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 2014, pp. 18–22, doi: 10.3115/v1/w14-2508.

[110] W. Y. Wang, “‘Liar, Liar Pants on Fire’: A New Benchmark Dataset for Fake News Detection,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, vol. 2, pp. 422–426, doi: 10.18653/v1/P17-2067.

[111] S. M. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, “SemEval 2016 Task 6 Detecting Stance in Tweets,” in *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval 2016*, 2016, pp. 31–41, doi: 10.18653/v1/s16-1003.