A Survey on Stance Detection for Mis- and Disinformation Identification

Anonymous ACL submission

Abstract

Understanding attitudes expressed in texts, also known as *stance detection*, plays an important role in systems for detecting false information online, be it misinformation (unintentionally false) or disinformation (intentionally false information). Stance detection has been framed in different ways, including (a) as a component of fact-checking, rumour detection, and detecting previously fact-checked claims, or (b) as a task in its own right. While there have been prior efforts to contrast stance detection with other related tasks such as argumentation mining and sentiment analysis, there is no existing survey on examining the relationship between stance detection and mis- and disinformation detection. Here, we aim to bridge this gap by reviewing and analysing existing work in this area, with mis- and disinformation in focus, and discussing lessons learnt and future challenges.

1 Introduction

The past decade is characterized by a rapid growth in popularity of social media platforms such as Facebook, Twitter, Reddit, and more recently, Parler. This, in turn, has led to a flood of dubious content, especially during controversial events such as Brexit and the US presidential election. More recently, with the emergence of the COVID-19 pandemic, social media were at the center of the first global infodemic (Alam et al., 2021), thus raising yet another red flag and a reminder of the need for effective mis- and disinformation detection online.

In this survey, we examine the relationship between automatically detecting false information online – including fact-checking, and detection of fake news, rumors, and hoaxes – and the core underlying Natural Language Processing (NLP) task needed to achieve this, namely stance detection. Therein, we consider mis- and disinformation, which both refer to false information, though disinformation has an additional intention to harm.

Detecting and aggregating the expressed stances towards a piece of information can be a powerful tool for a variety of tasks like understanding ideological debates (Hasan and Ng, 2014), gathering different frames of a particular issue (Shurafa et al., 2020) or determining the leanings of media outlets (Stefanov et al., 2020). The task of stance detection has been studied from different angles, e.g., in political debates (Habernal et al., 2018), for fact-checking (Thorne et al., 2018), or regarding new products (Somasundaran et al., 2009). Moreover, different types of text have been studied, including social media posts (Zubiaga et al., 2016b) and news articles (Pomerleau and Rao, 2017). Finally, stances expressed by different actors have been considered, such as politicians (Johnson et al., 2009), journalists (Hanselowski et al., 2019), and users on the web (Derczynski et al., 2017).

There are some recent surveys related to stance detection. Zubiaga et al. (2018a) discuss the role of stance in rumour verification, ALDayel and Magdy (2021) survey stance detection for social media, and Küçük and Can (2020) survey stance detection holistically, without a specific focus on veracity. There are also surveys on fact checking (Thorne and Vlachos, 2018; Guo et al., 2021), which mention though do not exhaustively survey stance.

However, there is no existing overview of the role different formulations of stance detection play in the detection of false content. In that respect, stance detection could be modelled as fact-checking — to gather stances of users or texts towards a claim or headline (to aid in fact checking or studying misinformation) —, or a component of a system that uses stance as part of its process of determining the veracity of an input claim. This paper aims to bridge this gap by surveying work on stance for mis- and disinformation detection, including task formulations, datasets, methods, from which we draw conclusions and lessons learnt, and a forecast of future research trends.

Dataset	Source(s)	Target	Context	Evidence	#Instances	Task
		English Datasets				
Rumour Has It (Qazvinian et al., 2011)	Twitter	Topic	Tweet	Multiple	10K	Rumours
PHEME (Zubiaga et al., 2016b)	Twitter	Claim	Tweet	Thread	4.5K	Rumours
Emergent (Ferreira and Vlachos, 2016)	News	Headline	Article*	Multiple	2.6K	Rumours
FNC-1 (Pomerleau and Rao, 2017)	News	Headline	Article	Single	75K	Fake news
RumourEval '17 (Derczynski et al., 2017)	Twitter	Implicit[1]	Tweet	Thread	7.1K	Rumours
FEVER (Thorne et al., 2018)	W	Claim	Facts	Multiple	185K	Fact-checking
Snopes (Hanselowski et al., 2019)	Snopes	Claim	Snippets	Multiple	19.5K	Fact-checking
RumourEval '19 (Gorrell et al., 2019)	Twitter, Reddit	Implicit[1]	Post	Thread	8.5K	Rumours
COVIDLies (Hossain et al., 2020)	Twitter	Claim	Tweet	Single	6.8K	Misconceptions
TabFact (Chen et al., 2020)	W	Statement	WikiTable	Multiple	118K	Fact-checking
		Non-English Datasets				
Arabic FC (Baly et al., 2018b)	News	Claim	Document	Single	3K	Fact-checking
DAST (Danish) (Lillie et al., 2019)	Reddit	Submission	Comment	Thread	3K	Rumour
Croatian (Bošnjak and Karan, 2019)	News	Title	Comment	Single	0.9K	Claim verifiability
ANS (Arabic) (Khouja, 2020)	News	Claim	Title	Single	3.8K	Claim verification
Ara(bic)Stance (Alhindi et al., 2021)	News	Claim	Title	Single	4K	Claim verification

Table 1: Key characteristics of stance detection datasets for mis- and disinformation detection. *#Instances* denotes dataset size as a whole; the numbers are in thousands (K) and are rounded to the hundreds. *the article's body is summarised. *Sources*: Twitter, News, Wikipedia, Reddit. *Evidence*: Single, Multiple, Thread.

2 Stance and Factuality

Here, we provide an overview of mis- and disinformation detection settings for which stance detection has been applied. As shown in Figure 2, stance can be used (a) as a way to perform fact-checking, or more typically (b) as a component of a fact-checking pipeline. Table 1 provides an overview of the key characteristics of available datasets. We include the *source* from which the data is collected and the *target*[1] towards which the stance is expressed in the provided *context*. We further show the type of evidence: *Single* is a single document/fact, *Multiple* is multiple pieces of textual evidence, often facts or documents, *Thread* is a (conversational) sequence of posts or a discussion. The final column is the type of the target *Task*. Finally, we present a dataset-agnostic summary of the terminology used for the different types of stance (see Figure 1), which we describe in a four-level taxonomy: (*i*) sources, i.e., where the dataset was collected from, (*ii*) inputs that represent the stance target (e.g., claim), and the accompanying context (e.g., news article), (*iii*) categorisation – meta-level characteristics of the input, and (*iv*) the textual object types for a particular stance scenario (e.g., topic, tweet, etc.). Appendix B discusses different stance scenarios with corresponding contexts and targets, with illustrations in Table 2.

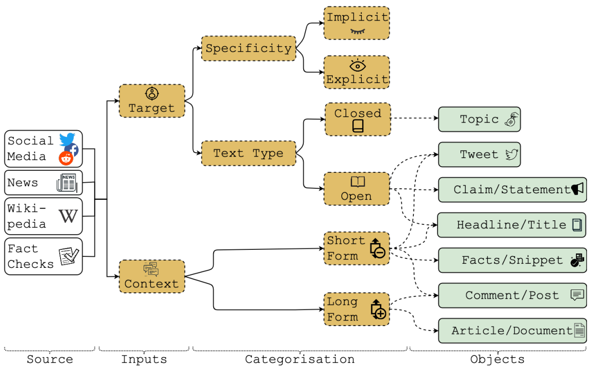


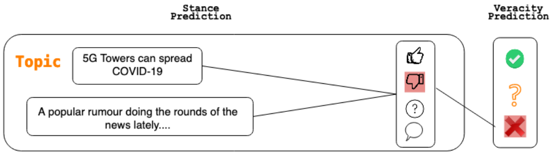
Figure 1: Types of stance. The *Target* is the object of the stance expressed in the *Context*.

2.1 Fact-Checking as Stance Detection

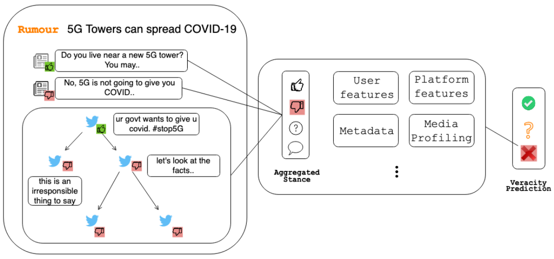
As stance detection is the core task within fact-checking, prior work has studied it in isolation, e.g., predicting the stance towards one or more documents. More precisely, the stance of the textual evidence(s) toward the target claim is considered as a veracity label, as illustrated in Figure 2a.

Fact-Checking with One Evidence Document Pomerleau and Rao (2017) organised the first Fake News Challenge (FNC-1) with the aim of automatically detecting fake news. The goal was to detect the relatedness of a news article's body w.r.t. a headline (possibly from another news article), based on the stance that the former takes regarding the latter. The possible categories were *positive*, *negative*, *discuss*, and *unrelated*. This was a standalone task, as it provides stance annotations only,

[1]The target can either be explicit, e.g., a topic such as *Public Healthcare*, or implicit, where only the context is present and the target is not directly available and is usually a topic (Derczynski et al., 2017; Gorrell et al., 2019), e.g., *Germanwings*, or '*Prince to play in Toronto*'. When the target is implicit, the task becomes similar to sentiment analysis.



(a) Stance detection as fact-checking



(b) Stance as a component of fact-checking pipeline

Figure 2: Two stance detection formulations.

omitting the actual "truth labels", with the motivation of assisting fact checkers in gathering several distinct arguments pertaining to a particular claim.

Fact-Checking with Multiple Evidence Documents The FEVER (Thorne et al., 2018, 2019) shared task was introduced in 2018 and extended in 2019, with the goal of determining the veracity of a claim based on a set of statements from Wikipedia. However, claims can be composite and can contain multiple (contradicting) statements, which requires multi-hop reasoning. The claim–evidence pairs are annotated as *SUPPORTED*, *REFUTED*, and *NOT ENOUGH INFO*. The latter category includes claims that are either too general or too specific, and therefore cannot be supported or refuted by the available information in Wikipedia. This setup may help fact checkers understand the decisions a model made in their assessment of the veracity of a claim, or assist human fact checkers.

The second edition (2019) of FEVER evaluated the robustness of models to adversarial attacks, where participants were tasked with providing new examples to "break" existing models, then propose "fixes" to improve system robustness to attacks. Note that FEVER slightly differs from typical stance detection, as it considers evidence supporting or refuting a claim, rather than the stance of an author towards a claim. An alternative way to look at this is in terms of argument reasoning, i.e., extracting and providing evidence for a claim. There is also a connection to Natural Language Inference, i.e. determining the relationship between sentence pairs. We still view FEVER as requiring stance detection as it resembles FNC, which is commonly seen as a stance detection task.

Apart from FEVER, Hanselowski et al. (2019) presented a task constructed from manually fact-checked claims on Snopes. For this task, a model had to predict the stance of evidence sentences in articles written by journalists towards claims. Unlike FEVER, this task does not require multi-hop reasoning. Chen et al. (2020) study the verification of claims using tabular data. The TabFact dataset was generated by human annotators who created positive and negative statements about Wikipedia tables. Two different forms of reasoning in a statement are required: (*i*) linguistic, i.e., semantic-level understanding, and (*ii*) symbolic, i.e., execution on the tables' structure.

2.2 Stance as a (Mis-/Dis-)information Detection Component

Fully automated systems can assist in gauging the extent, and studying the spread, of false information that propagates online. Hence, in contrast to the previously discussed applications of stance detection – as a stand-alone system for identification of mis- and disinformation, we here review its potency to serve as a component in a larger automated pipeline. Figure 2b shows an example of the setup, which can also include steps such as modelling the user, or profiling the media outlet among others. We discuss in more details the topics of media profiling, and misconceptions in Appendix C.

Rumors Stance detection can be used for rumour detection and debunking, where the stance of the crowd, media, or other sources towards a claim are used to determine the veracity of a currently circulating story or report of uncertain or doubtful factuality. More formally, *for a textual input and a rumour expressed as text, stance detection here is to determine the position of the text towards the rumour as a category label from the set Support, Deny, Query, Comment.* Zubiaga et al. (2016b) define these categories as whether the author: supports (*Support*) or denies (*Deny*) the veracity of the rumour they are responding to, "asks for additional evidence in relation to the veracity of the rumour" (*Query*) or "makes their own comment without a clear contribution to assessing the veracity of the rumour" (*Comment*). This setup has been widely explored for microblogs and social media. Qazvinian et al. (2011) started with five rumours and classified the user's stance as *endorse*, *deny*, *unrelated*, *question*, or *neutral*. While they are one of the first to demonstrate the feasibility of this task formulation; the limited size of their study and the focus on assessing stance of individual posts means their study has a limited real-world applicability.

Zubiaga et al. (2016b) analysed how people spread rumours on social media based on conversational threads. They included rumour threads associated with nine newsworthy events, and users' stance before and after the rumours were confirmed or denied. Dungs et al. (2018) continued this line of research, but focused on the effectiveness of stance for predicting rumour veracity. Hartmann et al. (2019) explored the flow of (dis-)information on Twitter after the MH17 Plane Crash.

The two RumourEval (Derczynski et al., 2017; Gorrell et al., 2019) shared tasks on automated claim validation aimed to identify and handle rumours based on user reactions and ensuing conversations in social media, offering annotations for both stance and veracity. The two editions of RumourEval were similar in spirit, with the second one providing more tweets and also additionally Reddit posts. RumourEval demonstrated the importance of modelling the context of a story instead of drawing conclusions based on a single post.

Ferreira and Vlachos (2016) debunked rumours based on news articles as part of the Emergent project. They collected claims and news articles from rumour sites with annotations both for stance and for veracity, done by journalists. The goal was to use the stance of a news article (summarised into a single sentence) towards a claim as one of the components to determine its veracity. A downside of this approach is the need of summarisation in contrast to FNC-1 (Pomerleau and Rao, 2017), where entire news articles were used.

Multiple languages All the above research has focused exclusively or primarily on English. Nevertheless, interest in stance detection for other languages has started to emerge. Baly et al. (2018b) integrated stance detection and fact-checking for Arabic in a single corpus. Khouja (2020) proposed a dataset for Arabic following the FEVER setup. Alhindi et al. (2021) introduced AraStance, a multi-country and multi-domain dataset of Arabic stance detection for fact-checking. Lillie et al. (2019) collected data for stance and veracity from Danish Reddit threads, annotated using the (S)upport, (D)eny, (Q)uery, (C)omment schema Zubiaga et al. (2016b). Bošnjak and Karan (2019) studied stance detection, claim verification, and sentiment analysis of comments for Croatian news articles.

3 Methods

In this section we discuss various ways to use stance detection for mis- and disinformation detection. We outline the state-of-the-art in Appendix D.

Fact Checking as Stance Detection Here, we discuss approaches for stance detection in the context of mis- and disinformation detection, where veracity is modelled as stance detection as outlined in Section 2.1. One such line of research is the Fake News Challenge. The competition organisers used weighted accuracy as an evaluation measure (FNC score), to mitigate the impact of class imbalance. Subsequently, Hanselowski et al. (2018a) criticized FNC score and F1-micro, and argued in favour of F1-macro (F1) instead. In the competition, most teams used models based on rich hand-crafted features such as words, word embeddings, and sentiment lexicons (Riedel et al., 2017; Hanselowski et al., 2018a). Hanselowski et al. (2018a) showed that the most important group of features were the lexical ones, followed by features from topic models, while sentiment analysis did not help. Ghanem et al. (2018) investigated the importance of lexical cue words, and found that *report* and *negation* are most beneficial, while *knowledge* and *denial* are least useful. All the above models struggle to learn the *Disagree* class, achieving up to 18 F1 due to major class imbalance. In contrast, *Unrelated* is detected almost perfectly by all models (over 99 F1). Hanselowski et al. (2018a) showed that these models exploit the lexical overlap between the headline and the document, but fail when there is a need to model semantic relations or complex negation, or to understand propositional content in general. This can be attributed to the use of n-grams, topic models, and lexicon-based features.

Mohtarami et al. (2018) investigated memory networks, aiming to mitigate the impact of irrelevant and noisy information by learning a similarity matrix and stance filtering component, and taking a step towards explaining the stance of a given claim by extracting meaningful snippets from evidence documents. However, their model also performs poorly on the *Agree/Disagree* classes due to the unsupervised way of learning memory networks for the task, i.e., there are no gold snippets justifying the document's stance w.r.t. the target claim.

More recently, transfer learning with pre-trained Transformer models has been explored (Slovikovskaya and Attardi, 2020), significantly surpassing results of previous approaches. Guderlei and Aßenmacher (2020) showed the most important hyper-parameter to be the learning rate, while freezing layers does not help. In particular, using a pre-trained Transformer such as RoBERTa, improves F1 for the *Disagree* (from 18 to 58) and the *Agree* (from 50 to 70) classes. The success of these models is also seen in cross-lingual settings. For Arabic, Khouja (2020) achieved 76.7 F1 on for stance detection on the ANS dataset using mBERT. Similarly, Hardalov et al. (2021b) applied pattern-exploiting training (PET) with sentiment pre-training in a cross-lingual setting showing sizeable improvements on 15 datasets. Alhindi et al. (2021) showed that language-specific pre-training was extremely important, also outperforming the state of the art on AraStance (52 F1) and Arabic FC (78 F1).

Some formulations include an extra step for evidence retrieval, e.g. retrieving Wikipedia snippets for FEVER (Thorne et al., 2018). To evaluate the whole fact checking pipeline, they introduced the FEVER score – the proportion of claims for which both correct evidence is returned and a correct label is predicted. The top systems that participated in the FEVER competition Hanselowski et al. (2018b); Yoneda et al. (2018); Nie et al. (2019) used LSTM-based models for natural language inference, e.g., enhanced sequential inference model (ESIM Chen et al. (2017)). Nie et al. (2019) proposed a neural semantic matching network, which ranked first in the competition, achieving 64.2 FEVER score. They used page view frequency and WordNet features in addition to pre-trained contextualized embeddings (Peters et al., 2018).

More recent approaches used bi-directional attention (Li et al., 2018), a GPT language model (Malon, 2018; Yang et al., 2019), and graph neural networks (Zhou et al., 2019; Atanasov et al., 2019; Liu et al., 2020; Wang et al., 2020; Zhong et al., 2020; Weinzierl et al., 2021; Si et al., 2021). Zhou et al. (2019) showed that adding graph networks on top of BERT can improve performance, reaching 67.1 FEVER score. Yet, the retrieval model is also important, e.g., using the gold evidence set adds 1.4 points. Liu et al. (2020); Zhong et al. (2020) replaced the retrieval model with a BERT-based one, in addition to using an improved mechanism to propagate the information between nodes in the graph, boosting the score to 70. Recently, Ye et al. (2020) experimented with a retriever that incorporates co-reference in distant-supervised pre-training (CorefRoBERTa). Wang et al. (2020) added external knowledge to build a contextualized semantic graph, setting a new SOTA on Snopes. Si et al. (2021) improved multi-hop reasoning using a model with eXtra Hop attention Zhao et al. (2020)), a capsule network aggregation layer, and LDA topic information.

Another notable idea is to use pre-trained language models as fact checkers based on a masked language modelling objective (Lee et al., 2020b), or to use the perplexity of the entire claim with respect to the target document (Lee et al., 2020a). Such models do not require a retrieval step, as they use the knowledge stored in language models. However, they are prone to biases in the patterns used, e.g., they can predict date instead of city/country and vice-versa when using “born in/on”. Moreover, the insufficient context can seriously confuse them, e.g., for short claims with uncommon words such as “Sarawak is a ...”, where it is hard to detect the entity type. Finally, the performance of such models remains well below supervised approaches; even though recent work shows that *few-shot training* can improve results (Lee et al., 2021).

Error analysis suggests the main challenges are (*i*) confusing the semantics at the sentence level, (*ii*) sensitivity to spelling errors, (*iii*) lack of relation between the article and the entities in the claim, (*vi*) dependence on syntactic overlaps, (*v*) embedding-level confusion, e.g., numbers tend to have similar embeddings, similarly for months.

Threaded Stance An alternative setting are conversational threads (Zubiaga et al., 2016b; Derczynski et al., 2017; Gorrell et al., 2019). In contrast to the single-task setup, which ignores or does not provide additional context, here, important knowledge can be gained from the structure of user interactions. These approaches are mostly applied as part of a larger system, e.g., for detecting and debunking rumours (see Section 2.2, *Rumours*). A common pattern is to use tree-like structured models, fed with lexicon-based content formatting (Zubiaga et al., 2016a) or dictionary-based token scores (Aker et al., 2017). Kumar and Carley (2019) replaced CRFs with Binarised Constituency Tree LSTMs, and used pre-trained embeddings to encode the tweets. More recently, Tree (Ma and Gao, 2020) and Hierarchical (Yu et al., 2020) Transformers were proposed which combine post- and thread-level representations for rumour debunking, improving previous results on RumourEval ’17 (Yu et al., 2020). Kochkina et al. (2017, 2018) split conversations into branches, modelling each branch with branched-LSTM and hand-crafted features, outperforming other systems at RumourEval '17 on stance detection (43.4 F1). Li et al. (2020) deviated from this structure and modelled the conversations as a graph. Tian et al. (2020) showed that pre-training on stance data yielded better representations for threaded tweets for downstream rumour detection. Yang et al. (2019) curated per-class pre-training data by adapting examples, not only from stance datasets, but also from tasks such as question answering achieving the highest F1 (57.9) on the RumourEval '19 stance detection task. Li et al. (2019a,b) additionally incorporated user credibility information, conversation structure, and other content-related features ranking 3rd on stance detection and 1st on veracity classification (RumourEval '19). Finally, the stance of a post might not be expressed directly towards the root of the thread, thus the preceding posts must be also taken into account (Gorrell et al., 2019).

A major challenge for all rumour detection datasets is the class distribution, e.g., the minority class *denying* is extremely hard for models to learn, as even for strong systems such as Kochkina et al. (2017) the F1 for it is 0. Label semantics also appears to play a role as the *querying* label has a similar distribution, but much higher F1. Yet another factor is thread depth, as performance drops significant at higher depth, especially for the *supporting* class. On the positive side, using multi-task learning and incorporating stance detection labels into veracity detection yields a huge boost in performance (Gorrell et al., 2019; Yu et al., 2020).

Another factor is the temporal dimension of posts in a thread (Lukasik et al., 2016; Veyseh et al., 2017; Dungs et al., 2018; Wei et al., 2019). In-depth data analysis (Zubiaga et al. (2016a,b); Kochkina et al. (2017); Wei et al. (2019); Ma and Gao (2020); Li et al. (2020); among others) shows interesting patterns along the temporal dimension: (*i*) source tweets (at zero depth) usually support the rumour and models often learn to detect that, (*ii*) it takes time for denying tweets to emerge, afterwards for false rumors their number increases quite substantially, (*iii*) the proportion of querying tweets towards unverified rumors also shows an upward trend over time, but their overall number decreases.

Multi-Dataset Learning (MDL) Mixing data from different domains and sources can improve the robustness of models. However, setups that combine mis- and disinformation identification with stance detection, outlined in Section 2, vary in their annotation and labelling schemes, which poses many challenges.

Earlier approaches focused only on the pre-training of models on multiple tasks, e.g., Fang et al. (2019) achieved state-of-the-art results on FNC-1 by fine-tuning on multiple tasks such as question answering, natural language inference, etc., which are weakly related to stance detection. Recently, Schiller et al. (2021) proposed a stance detection benchmark to evaluate the robustness of stance models. They leveraged a pre-trained multi-task deep neural network (MT-DNN Liu et al. (2019)) and continued its training on all datasets simultaneously using multi-task learning, showing sizeable improvements over strong baselines trained on individual datasets. Hardalov et al. (2021a) explored the possibility of cross-domain learning from sixteen stance detection datasets. They proposed a novel architecture (MoLE), which combines domain adaptation techniques applied at different stages of the modelling process (Luo et al., 2002) – feature-level (Guo et al., 2018; Wright and Augenstein, 2020) and decision-level (Ganin and Lempitsky, 2015). They further integrated label embeddings (Augenstein et al., 2018), and eventually developed an end-to-end unsupervised framework for predicting stance from a set of unseen target labels (which are out-of-domain). Hardalov et al. (2021b) explored PET (Schick and Schütze, 2021) for cross-lingual setting, combining datasets with different label inventories. They do so by modelling the task as a cloze question answering one, showing that MDL helps somewhat for low-resource and substantively for full-resource scenarios. Moreover, transferring knowledge from English stance datasets and noisily generated sentiment-based stance data can further boost performance.

4 Lessons Learnt and Future Trends

Dataset Size A major limitation holding back the performance of machine learning based stance detection is the size of existing stance datasets, the vast majority of which contain at most a few thousand examples. Contrasted with the related task of Natural Language Inference, where datasets such as SNLI (Bowman et al., 2015) of more than half a million samples have been collected, this is far from optimal. Moreover, the small dataset sizes are often accompanied with skewed class distribution with very few examples from the minority classes, including many of the datasets in this study (Zubiaga et al., 2016b; Derczynski et al., 2017; Pomerleau and Rao, 2017; Baly et al., 2018b; Gorrell et al., 2019; Lillie et al., 2019; Alhindi et al., 2021). This can lead to a significant disparity for label performance as outlined in Section 3. Several techniques have been proposed for mitigating this, such as sampling strategies (Nie et al., 2019), weighting classes (Veyseh et al., 2017),[2] crafting artificial examples from auxiliary tasks (Yang et al., 2019; Hardalov et al., 2021b), or training on multiple datsets (Schiller et al., 2021; Hardalov et al., 2021a,b).

Data Mixing A potential way of overcoming the resource limitation and narrow focus of the data is to combine several datasets. Yet, as we previously discussed (see Section 2), task definitions and label inventories vary across stance datasets. Further, large-scale studies of approaches that leverage the relationships between label inventories, or the similarity between datasets are still largely lacking. One promising direction is the use of label embeddings (Augenstein et al., 2018), as they offer a convenient way to learn interactions between disjoint label sets that carry semantic relations. One such first study was recently presented by Hardalov et al. (2021a), which explored different strategies for leveraging inter-dataset signals and label interactions in both in- (seen targets) and out-of-domain (unseen targets) settings. This could help to overcome challenges faced by models trained on small-size datasets, and even smaller minority classes.

Multilinguality Multi-linguality is important for several reasons: (*i*) the content may originate in various languages, (*ii*) the evidence or the stance may not be expressed in the same language, thus (*iii*) posing a challenge for fact-checkers, who might not be speakers of the language the claim was originally made in, and (*iv*) it adds more data that can be leveraged for modelling stance. Currently, only a handful of datasets for factuality and stance cover languages other than English (see Table 1), and they are small in size and do not offer a cross-lingual setup. Recently, Vamvas and Sennrich (2020) proposed such a setup for three languages for stance in debates, Schick and Schütze (2021) explored few-shot learning, and Hardalov et al. (2021b) extended that paradigm with sentiment and stance pre-training and evaluated on twelve languages from various domains.

[2]Weighting is not trivial for some setups, e.g., threaded stance (Zubiaga et al., 2018b)

Since cultural norms and expressed linguistic phenomena play a crucial role in understanding the context of a claim (Sap et al., 2019), we do not argue for a completely language-agnostic framework. Yet, empirically, training in cross-lingual setups helps improve performance by leveraging better representations by training on a similar language or by acting as a regulariser.

Modelling Context Modelling context is a particularly important, yet challenging task. In many cases, there is a need to consider the background of the stance-taker as well as the characteristics of the targeted object. In particular, in the context of social media, one can provide information about the users such as their previous activity, other users they interact most with, the threads in which they participate, or even their interests (Zubiaga et al., 2016b; Gorrell et al., 2019; Li et al., 2019b). The context of the stance expressed in news articles is related to the features of the media outlets, such as source of funding, previously known biases, or credibility (Baly et al., 2019a; Darwish et al., 2020; Stefanov et al., 2020; Baly et al., 2020). When using contextual information about the object, factual information about the real world, and the time of posting are all important. Incorporating those into a stance detection pipeline, while challenging, paves the way towards a robust detection process.

Multimodal Content Spreading mis- and disinformation through multiple modalities is becoming increasingly popular. One such example are *DeepFakes*, i.e., synthetically created images or videos, in which (usually) the face of one person is replaced with another person's face. Another such example are information propagation techniques such as *memetic warfare*, i.e., the use of memes for information warfare. Hence it is increasingly important to combine different modalities to understand the full context the stance is being expressed in. Some work in this area is on fake news detection for images (Nakamura et al., 2020), claim verification for images (Zlatkova et al., 2019), or searching for fact-checked information to alleviate the spread of fake news (Vo and Lee, 2020). There has been work on meme analysis for related tasks: Hateful Memes Challenge (Kiela et al., 2020) and SemEval-2021 Task 6 on Detection of Persuasion Techniques in Texts and Images (Dimitrov et al., 2021). This line of research is especially relevant for mis- and disinformation tasks that depend on the wisdom of the crowd as expressed on social media (e.g., Twitter or Reddit) as it adds additional information sources (Qazvinian et al., 2011; Zubiaga et al., 2016b; Derczynski et al., 2017; Hossain et al., 2020); see Section 4, *Modelling context*.

Shades of Truth The notion of *shades of truth* is important in mis- and disinformation detection. For example, fact checking often goes beyond binary *true*/*false* labels, e.g., Nakov et al. (2018) used a third category *half-true*, Rashkin et al. (2017) included *mixed* and *no factual evidence*, and Wang (2017); Santia and Williams (2018) adopted an even finer-grained schema with six labels, including *barely true* and *utterly false*. We believe that such shades could be applied to stance and used in a larger pipeline. In fact, fine-grained labels are common for the related task of Sentiment Analysis (Pang and Lee, 2005; Rosenthal et al., 2017).

Label Semantics As research in stance detection has evolved, so has the definition of the task and the label inventories, however they still do not capture the strength of the expressed stance. As shown in Section 2 (also Appendix A), labels can vary based on the use case and the setting they are used in. Most researchers have adopted a variant of the *Favour*, *Against*, and *Neither* labels, or an extended schema such as *(S)upport*, *(Q)uery*, *(D)eny*, and *(C)omment* (Mohammad et al., 2016), but that is not enough to accurately assess stance. Furthermore, adding label granularity can further improve transfer among dataset, as the stance labels already share some semantic similarities, however there can be mismatches in the label definitions (Schiller et al., 2021; Hardalov et al., 2021a,b).

Explainability The ability to explain model decisions is important, especially for mis- and disinformation detection, as one could argue it is a crucial step towards adopting fully automated fact checking. FEVER 2.0 (Thorne et al., 2019) may be viewed as a step towards obtaining such explanations, e.g., there have been efforts to identify adversarial triggers that offer explanations for the vulnerabilities at the model level (Atanasova et al., 2020b). However, FEVER is artificially created and is limited to Wikipedia, which may not reflect real-world settings. To mitigate this, explanation by professional journalists can be found on fact checking websites, and can be further combined with stance detection in an automated system. A step in this direction is Atanasova et al. (2020a), who generated natural language explanations for claims from PolitiFact given gold evidence document summaries by journalists. Moreover, partial explanations can be obtained automatically from the underlying models, e.g., from memory networks (Mohtarami et al., 2018), attention weights (Zhou et al., 2019; Liu et al., 2020), or topic relations (Si et al., 2021). However, such approaches are limited as they can require gold snippets justifying the document's stance, attention weights can be misleading (Jain and Wallace, 2019), and topics might be noisy due to their unsupervised nature. Other existing systems (Popat et al., 2017, 2018; Nadeem et al., 2019) offer explanations to a more limited extent, highlighting span overlaps between the target text and the evidence documents. Overall, there is a need for holistic and realistic explanations of how a fact checking model arrived at its prediction.

Integration People question false information more and tend to confirm true information (Mendoza et al., 2010). Thus, stance can play a vital role in verifying dubious content. In Appendix E, we discuss existing systems and real-world applications of stance for mis- and disinformation identification in more detail. However, we argue that a tighter integration between stance and fact checking is needed. Stance can be expressed in different forms, e.g., tweets, news articles, user posts, sentences in Wikipedia, and Wiki tables, among others and can have different formulations as part of the fact-checking pipeline (see Section 2). All these can guide human fact checkers through the process of fact checking, and can point them to relevant evidence. Moreover, the wisdom of the crowd can be a powerful instrument in the fight against mis- and disinformation (Pennycook and Rand, 2019), but we should note that vocal minorities can derail public discourse (Scannell et al., 2021). These risks can be mitigated by taking into account the credibility of the user or of the information source.

5 Conclusion

We surveyed the current state-of-the-art in stance detection for mis- and disinformation detection. We explored applications of stance for detecting fake news, verifying rumours, identifying misconceptions, and fact checking. We also discussed existing approaches used in different aspects of the aforementioned tasks, and we outlined several interesting phenomena, which we summarised as lessons learned and promising future trends.

References

Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. 2017. Simple open stance classification for rumour analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 31–39, Varna, Bulgaria. INCOMA Ltd.

Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, and Preslav Nakov. 2021. Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):913–922.

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Tariq Alhindi, Amal Alabdulkarim, Ali Alshehri, Muhammad Abdul-Mageed, and Preslav Nakov. 2021. AraStance: A multi-country and multi-domain dataset of Arabic stance detection for fact checking. In *Proceedings of the 4th Workshop on NLP for Internet Freedom: Censorship, Disinformation and Propaganda*, NLP4IF '21.

Atanas Atanasov, Gianmarco De Francisci Morales, and Preslav Nakov. 2019. Predicting the role of political trolls in social media. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1023–1034, Hong Kong, China. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020a. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020b. Generating label cohesive and well-formed adversarial claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177, Online. Association for Computational Linguistics.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. 2018. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1896–1906, New Orleans, Louisiana. Association for Computational Linguistics.

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

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3364–3374, Online. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019a. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2109–2116, Minneapolis, Minnesota. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019b. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2109–2116, Minneapolis, Minnesota. Association for Computational Linguistics.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018b. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, Louisiana. Association for Computational Linguistics.

Douglas Biber and Edward Finegan. 1988. Adverbial stance types in English. *Discourse Processes*, 11(1):1–34.

Mihaela Bošnjak and Mladen Karan. 2019. Data set for stance and sentiment analysis from user comments on Croatian news. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 50–55, Florence, Italy. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 141–152.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 Task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21.

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, 8(9):938–949.

John W Du Bois. 2007. The stance triangle. *Stancetaking in discourse: Subjectivity, evaluation, interaction*, 164(3):139–182.

Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can rumour stance alone predict veracity? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3360–3370, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Wei Fang, Moin Nadeem, Mitra Mohtarami, and James Glass. 2019. Neural multi-task learning for stance prediction. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 13–19, Hong Kong, China. Association for Computational Linguistics.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France.

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

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

Maike Guderlei and Matthias Aßenmacher. 2020. Evaluating unsupervised representation learning for detecting stances of fake news. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6339–6349, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703, Brussels, Belgium. Association for Computational Linguistics.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2021. A survey on automated fact-checking. *arXiv preprint arXiv:2108.11896*.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

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

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018b. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021a. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021b. Few-shot cross-lingual stance detection with sentiment-based pre-training. *arXiv preprint arXiv:2109.06050*.

Mareike Hartmann, Yevgeniy Golovchenko, and Isabelle Augenstein. 2019. Mapping (dis-)information flow about the MH17 plane crash. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 45–55, Hong Kong, China. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Austin Hounsel, Jordan Holland, Ben Kaiser, Kevin Borgolte, Nick Feamster, and Jonathan Mayer. 2020. Identifying disinformation websites using infrastructure features. In *Proceedings of the 10th USENIX Workshop on Free and Open Communications on the Internet*, FOCI '20.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Jude Khouja. 2020. Stance prediction and claim verification: An Arabic perspective. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 8–17, Online. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, NeurIPS '20.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480, Vancouver, Canada. Association for Computational Linguistics.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Sumeet Kumar and Kathleen Carley. 2019. Tree LSTMs with convolution units to predict stance and rumor veracity in social media conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5047–5058, Florence, Italy. Association for Computational Linguistics.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2020a. Misinformation has high perplexity. *arXiv preprint arXiv:2006.04666*.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1971–1981, Online. Association for Computational Linguistics.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020b. Language models as fact checkers? In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41, Online. Association for Computational Linguistics.

Jiawen Li, Yudianto Sujana, and Hung-Yu Kao. 2020. Exploiting microblog conversation structures to detect rumors. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5420–5429, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019a. 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*, pages 855–859, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019b. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1173–1179, Florence, Italy. Association for Computational Linguistics.

Sizhen Li, Shuai Zhao, Bo Cheng, and Hao Yang. 2018. An end-to-end multi-task learning model for fact checking. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 138–144, Brussels, Belgium. Association for Computational Linguistics.

Anders Edelbo Lillie, Emil Refsgaard Middelboe, and Leon Derczynski. 2019. Joint rumour stance and veracity prediction. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 208–221, Turku, Finland. Linköping University Electronic Press.

Qian Liu, Bei Chen, Jiaqi Guo, Zeqi Lin, and Jianguang Lou. 2021. TAPEX: Table pre-training via learning a neural SQL executor. *arXiv preprint arXiv:2107.07653*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Michal Lukasik, P. K. Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. 2016. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 393–398, Berlin, Germany. Association for Computational Linguistics.

R.C. Luo, Chih-Chen Yih, and Kuo Lan Su. 2002. Multisensor fusion and integration: approaches, applications, and future research directions. *IEEE Sensors Journal*, 2(2):107–119.

Jing Ma and Wei Gao. 2020. Debunking rumors on Twitter with tree transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5455–5466, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Christopher Malon. 2018. Team papelo: Transformer networks at FEVER. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 109–113, Brussels, Belgium. Association for Computational Linguistics.

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter under crisis: Can we trust what we RT? In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, page 71–79, New York, NY, USA. Association for Computing Machinery.

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

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 767–776, New Orleans, Louisiana. Association for Computational Linguistics.

Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 353–362. ACM.

Moin Nadeem, Wei Fang, Brian Xu, Mitra Mohtarami, and James Glass. 2019. FAKTA: An automatic end-to-end fact checking system. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 78–83, Minneapolis, Minnesota. Association for Computational Linguistics.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6149–6157, Marseille, France. European Language Resources Association.

Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 372–387, Cham. Springer International Publishing.

An T. Nguyen, Aditya Kharosekar, Matthew Lease, and Byron C. Wallace. 2018. An interpretable joint graphical model for fact-checking from crowds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI '18, pages 1511–1518.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: leveraging social context for fake news detection using graph representation. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1165–1174. ACM.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *The Thirty-Third Conference on Artificial Intelligence AAAI 2019*, pages 6859–6866, Honolulu, Hawaii, USA. AAAI Press.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Gordon Pennycook and David G. Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7):2521–2526.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge stage 1 (FNC-I): Stance detection. https://www.fakenewschallenge.org/.

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

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018. CredEye: A credibility lens for analyzing and explaining misinformation. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 155–158, Lyon, France. International World Wide Web Conferences Steering Committee.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

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

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Giovanni C Santia and Jake Ryland Williams. 2018. Buzzface: A news veracity dataset with facebook user commentary and egos. In *Twelfth International AAAI Conference on Web and Social Media*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Denise Scannell, Linda Desens, Marie Guadagno, Yolande Tra, Emily Acker, Kate Sheridan, Margo Rosner, Jennifer Mathieu, and Mike Fulk. 2021. COVID-19 vaccine discourse on Twitter: A content analysis of persuasion techniques, sentiment and mis/disinformation. *Journal of Health Communication*, 26(7):443–459.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI-Künstliche Intelligenz*, pages 1–13.

Chereen Shurafa, Kareem Darwish, and Wajdi Zaghouani. 2020. Political framing: US COVID19 blame game. In *Social Informatics*, pages 333–351. Springer International Publishing.

Jiasheng Si, Deyu Zhou, Tongzhe Li, Xingyu Shi, and Yulan He. 2021. Topic-aware evidence reasoning and stance-aware aggregation for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1612–1622, Online. Association for Computational Linguistics.

Valeriya Slovikovskaya and Giuseppe Attardi. 2020. Transfer learning from transformers to fake news challenge stance detection (FNC-1) task. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1211–1218, Marseille, France. European Language Resources Association.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 527–537, Online. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. 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, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.

Lin Tian, Xiuzhen Zhang, Yan Wang, and Huan Liu. 2020. Early detection of rumours on twitter via stance transfer learning. In *Advances in Information Retrieval*, ECIR '20, pages 575–588. Springer International Publishing.

Jannis Vamvas and Rico Sennrich. 2020. X-Stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*.

Amir Pouran Ben Veyseh, Javid Ebrahimi, Dejing Dou, and Daniel Lowd. 2017. A temporal attentional model for rumor stance classification. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 2335–2338. ACM.

Nguyen Vo and Kyumin Lee. 2020. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7717–7731, Online. Association for Computational Linguistics.

William Yang Wang. 2017. "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)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yongyue Wang, Chunhe Xia, Chengxiang Si, Beitong Yao, and Tianbo Wang. 2020. Robust reasoning over heterogeneous textual information for fact verification. *IEEE Access*, 8:157140–157150.

Penghui Wei, Nan Xu, and Wenji Mao. 2019. Modeling conversation structure and temporal dynamics for jointly predicting rumor stance and veracity. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4787–4798, Hong Kong, China. Association for Computational Linguistics.

Maxwell Weinzierl, Suellen Hopfer, and Sanda M Harabagiu. 2021. Misinformation adoption or rejection in the era of COVID-19. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 787–795.

Weiming Wen, Songwen Su, and Zhou Yu. 2018. Cross-lingual cross-platform rumor verification pivoting on multimedia content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3487–3496, Brussels, Belgium. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2020. Transformer based multi-source domain adaptation. In

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7963–7974, Online. Association for Computational Linguistics.

Ruoyao Yang, Wanying Xie, Chunhua Liu, and Dong Yu. 2019. BLCU_NLP at SemEval-2019 task 7: An inference chain-based GPT model for rumour evaluation. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1090–1096, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186, Online. Association for Computational Linguistics.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. UCL machine reading group: Four factor framework for fact finding (HexaF). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102, Brussels, Belgium. Association for Computational Linguistics.

Jianfei Yu, Jing Jiang, Ling Min Serena Khoo, Hai Leong Chieu, and Rui Xia. 2020. Coupled hierarchical transformer for stance-aware rumor verification in social media conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1392–1401, Online. Association for Computational Linguistics.

Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul Bennett, and Saurabh Tiwary. 2020. Transformer-xh: Multi-evidence reasoning with extra hop attention. In *International Conference on Learning Representations*.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Online. Association for Computational Linguistics.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108, Hong Kong, China. Association for Computational Linguistics.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018a. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2).

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016a. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2438–2448, Osaka, Japan. The COLING 2016 Organizing Committee.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018b. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management*, 54(2):273–290.

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

A What is Stance?

In order to understand the task of stance detection, we first provide definitions of stance and the stance-taking process. Biber and Finegan (1988) define stance as the expression of a speaker's standpoint and judgement towards a given proposition. Further, Du Bois (2007)) define stance as "*A public act by a social actor, achieved dialogically through overt communicative means, of simultaneously evaluating objects, positioning subjects (self and others), and aligning with other subjects, with respect to any salient dimension of the sociocultural field*", showing that the stance-taking process is affected not only by one's personal opinion, but also by other external factors such as cultural norms, roles in the institution of the family, etc. Here, we adopt the general definition for stance detection from Küçük and Can (2020): "*for an input in the form of a piece of text and a target pair, stance detection is a classification problem where the stance of the author of the text is sought in the form of a category label from this set: Favor, Against, Neither. Occasionally, the category label of Neutral is also added to the set of stance categories (Mohammad et al., 2016), and the target may or may not be explicitly mentioned in the text* (Augenstein et al., 2016; Mohammad et al., 2016). Note that the stance detection definitions and the label inventories vary somewhat, depending on the target application (see Section 2).

Finally, stance detection can be distinguished from several other closely related NLP tasks: (*i*) *biased language detection*, where the existence of an inclination or tendency towards a particular perspective within a text is explored; (*ii*) *emotion recognition*, where the goal is to recognise emotions such as *love, anger, sadness, etc.* in the text; (*iii*) *perspective identification*, which aims to find the point-of-view of the author (e.g., Democrat vs. Republican) and the target is always explicit; (*iv*) *sarcasm detection*, where the interest is in satirical or ironic pieces of text, which are often written with the intent of ridicule or mockery; (*v*) *sentiment analysis*, which determines the polarity of a piece of text.

B Examples of Stance

As outlined in Section 2, there are different formulations in which the task of stance definition is materialised. In Table 2, we present some instances of these as exemplified by different stance detection datasets. The topic towards which the stance is assessed can vary e.g. *Headline*, *Comment*, *Claim*, *Topic* etc., which differ in length and form making modelling the task difficult. Further, the context where the stance is expressed can vary in not only in its domain (e.g., *News* in Ferreira and Vlachos (2016) and *Twitter* in Qazvinian et al. (2011)), but also in its structure, as seen in the example of multiple evidence sentences from Thorne et al. (2018) and threaded comments from Gorrell et al. (2019).

In a more detailed view of Table 2, we see that each group of examples has its own important specifics that alter the task of stance detection for mis- and disinformation detection.

Figure 2a shows an example from the *News* domain, where we have a headline and a whole article body, and the goal is to find how are the two related in terms of the body's stance(s) towards the headline. In this scenario, the models need to be able to handle very long documents, on one hand, and on the other to reason over multiple pieces of the text, that might potentially express different stances. It is possible to simplify that task by extracting a summary of the news article, beforehand, and evaluating only its stance, as shown in Figure 2d. Nonetheless, obtaining these summaries is not a trivial task, either they need to be extracted by a human annotator (e.g., journalist), which is time consuming and can be expensive, but also can require apriori knowledge for the headline/topic of interest, as the article might have more than one highlight (or viewpoint), another possibility for obtaining the summary can be machine summarisation, which can be noisy, and prone to errors.

Stances oftentimes are expressed in social media websites such as Twitter, Facebook, Reddit, etc. We illustrate two such scenarios in Figures 2b and 2e. In contrast to the usually long and well-written news documents, social media posts are mostly short in length, and depend on additional context such as the previous posts in a conversational thread (Figure 2e), or external URLs and implicit topics (Figure 2b). Moreover, these texts also need normalisation, as users tend to use slurs, emojis and other types of informal language.

Next, in Figure 2c we highlight another interesting setup – claim verification using multiple evidences. Here, the reasoning is carried in multiple hops over a set of texts. In particular, there might

[2] For illustrative purposes the text is trimmed to include only the relevant passage.

Headline: *Robert Plant Ripped up $800M Led Zeppelin Reunion Contract*
Body: ...Led Zeppelin's Robert Plant turned down £500 MILLION to reform supergroup..

(a) Example from Pomerleau and Rao (2017)

Topic: *Sarah Palin getting divorced?*
Tweet: OneRiot.com - Palin Denies First Dude Divorce Rumors http://url

Topic: *N/A (Implicit)*
Tweet: Wow, that is fascinating! I hope you never mock our proud Scandi heritage again.

(b) Examples from Qazvinian et al. (2011) and Derczynski et al. (2017)

Claim: *The Rodney King riots took place in the most populous county in the USA.*
Wiki **Evidence 1**: The 1992 Los Angeles riots, *also known as the Rodney King riots* were a series of riots, lootings, arsons, and civil disturbances that *occurred in Los Angeles County*, California in April and May 1992.
Wiki **Evidence 2**: Los Angeles County, officially the County of Los Angeles, *is the most populous county in the USA*.

(c) Example from Thorne et al. (2018)

Headline: *Jess Smith of Chatham, Kent was the smiling sun baby in the Teletubbies TV show*
Summary 1: Canterbury Christ Church University student Jess Smith, from Chatham, starred as Teletubbies sun
Summary 2: This College Student Claims She Was The Teletubbies Sun Baby

(d) Example from Ferreira and Vlachos (2016)

u1: We understand that there are two gunmen and up to a dozen hostages inside the cafe under siege at Sydney.. ISIS flags remain on display #7News
- **u2**: @u1 not ISIS flags
- **u3**: @u1 sorry - how do you know its an ISIS flag? Can you actually confirm that?
  - **u4**: @u3 no she cant cos its actually not
- **u5**: @u1 More on situation at Martin Place in Sydney, AU LINK
- **u6**: @u1 Have you actually confirmed its an ISIS flag or are you talking shit

(e) Example from Gorrell et al. (2019)

Table 2: Illustrative examples for different stance detection scenarios included in our survey. We annotate the expressed stance with (*support, for*), (*deny, against*), (*query*), and (*comment*).

not exists a single passage from a document/post that supports/refutes the claim, directly. In that case a large enough chain of evidence is needed, that can cover a sufficient amount of contextual knowledge, for can allow the model (or a person) to assess the veracity of a given claim.

Finally, the examples in Figure 2 demonstrate that stance can be used for mis- and disinformation information in different ways: (*i*) directly, as in the examples in Figures 2a and 2b, or (*ii*) as multiple viewpoints, which are later aggregated into a final decision label, Figure 2c, 2d and 2e.

We thoroughly discuss all of the aforementioned setups in Section 2, including the publicly available datasets that focus on stance in the context of mis- and disinformation identification.

C Additional Formulations of Stance as a Component

Beyond the approaches outlined in Section 2.2, stance has also been used in detecting misconceptions and profiling of media sources as part of the fact-checking pipeline. We describe work following those formulations here.

Misconceptions Hossain et al. (2020) focused on detecting misinformation related to COVID-19, based on a set of known misconceptions listed in Wikipedia. In particular, they evaluated the veracity of a tweet depending on whether it *agrees*, *disagrees*, or has *no stance* with respect to a subset of misconceptions most relevant to it. This may allow fact-checkers to assess the veracity of dubious content in a convenient way by evaluating the stance of a claim regarding already checked stories, known misconceptions, and facts.

Media profiling Stance detection has been also used for media profiling. Stefanov et al. (2020) explored the feasibility of an unsupervised approach for identifying the political leanings (left, center, or right bias) of media outlets and influential people on Twitter based on their stance on controversial topics. They built clusters of users around core vocal ones based on their behaviour on Twitter such as retweeting, using the procedure proposed by Darwish et al. (2020). This is an important step towards understanding media biases.

The reliability of entire news media sources has been automatically estimated based on their stance with respect to known manually fact-checked claims, without access to gold labels for the overall medium-level factuality of reporting (Mukherjee and Weikum, 2015; Popat et al., 2017, 2018). The assumption in such methods is that reliable media agree with true claims and disagree with false ones, while for unreliable media, the situation is reversed. The trustworthiness of Web sources has also been studied from a Data Analytics perspective. For instance, Dong et al. (2015) proposed that a trustworthy source is one that contains very few false claims.

More recently, Baly et al. (2018a) used gold labels from Media Bias/Fact Check,[3] and a variety of information sources: articles published by the medium, what is said about the medium on Wikipedia, metadata from its Twitter profile, URL structure, and traffic information. In follow-up work, (Baly et al., 2019b) used the same representation to jointly predict a medium's factuality of reporting (*high* vs. *mixed* vs. *low*) and its bias (*left* vs. *center* vs. *right*) on an ordinal scale, in a multi-task ordinal regression setup. Baly et al. (2020) extended the information sources to include Facebook followers and speech signals from the news medium's channel on YouTube (if any). Finally, Hounsel et al. (2020) proposed to use domain, certificate, and hosting information about the infrastructure of the hosting website.

D State-of-the-art

Table 3 lists the state-of-the-art (SOTA) results for each dataset discussed in Section 2 and Table 1. The datasets vary in the task formulation and in their composition in terms of size, number of classes, class imbalance, topics, metrics, etc. All of these factors impact performance, leading to sizable differences in the final score, as discussed Section 3, and hence rendering them not directly comparable to one another.

Paper	**Dataset**	**Score**	**Metric**
Hardalov et al. (2021a)	Rumour Has It	71.2	$F1_{macro}$
Kumar et al. (2019)	PHEME	53.2	$F1_{macro}$
Hardalov et al. (2021a)	Emergent	86.2	$F1_{macro}$
Guderlei et al. (2020)	FNC-1	78.2	$F1_{macro}$
Yu et al. (2020)	RumourEval '17	50.9	$F1_{macro}$
Dominiks (2021)*	FEVER	76.8	FEVER
Wang et al. (2020)	Snopes	78.3	$F1_{macro}$
Yang et al. (2019)	RumourEval '19	61.9	$F1_{macro}$
Weinzierl et al. (2021)	COVIDLies	74.3	$F1_{macro}$
Liu et al. (2021)	TabFact	84.2	Accuracy
Alhindi et al. (2021)	Arabic FC	52.?	$F1_{macro}$
Lillie et al. (2019)	DAST	42.1	$F1_{macro}$
Bošnjak and Karan (2019)	Croatian	25.8	$F1_{macro}$
Alhindi et al. (2021)	ANS	90.?	$F1_{macro}$
Alhindi et al. (2021)	AraStance	78.?	$F1_{macro}$

Table 3: State-of-the-art results on the stance detection datasets. Note that some papers round their results to integers, and thus we put '**?**' for them. *Extracted from the FEVER leaderboard.[4]

E Systems and Applications

The systems and applications below use stance detection as part of a pipeline for identifying mis- and disinformation, see Section 3 for more details about the methods.

Popat et al. (2018) proposed CredEye, a system for automatic credibility assessment of textual claims. It takes a claim as an input and analyses its credibility by considering relevant articles from the Web, by combining the predicted stance of the articles regarding the claim with linguistic features to obtain a credibility score (Popat et al., 2017).

Nguyen et al. (2018) designed a prototype fact-checker Web tool[5]. Their system leverages a probabilistic graphical model to assess a claim's veracity taking into consideration the stance of multiple articles regarding this claim, the reputation of the news sources, and the annotators' reliability. In addition, it offers explanations to the fact-checkers based on the aforementioned features, which was shown to improve the overall user satisfaction and trust in the predictions.

Zubiaga et al. (2018a) considered a four-step tracking process as a pipeline for rumour resolution: (*1*) *rumour detection*, which, given a stream of claims, determines whether they are worth verifying or they do not contain a rumour; (*2*) *rumour tracking* for finding relevant information about the rumour using social media posts, sentence descriptions, and keywords; (*3*) *stance classification* to collect stances towards that rumour; and (*4*) *veracity classification* to aggregate the information from the tracking component, the collected stances, and optionally other relevant information about sources, metadata about the users, etc., to predict a truth value for the rumour. Possible methods that can be applied at each step in the pipeline were also discussed in more detail.

[3] http://mediabiasfactcheck.com

[4] The result from *dominiks* can be found at https://competitions.codalab.org/competitions/18814#results

[5] http://fcweb.pythonanywhere.com/

Wen et al. (2018) worked in a cross-lingual cross-platform rumour verification setup. They included multimodal content from fake and from real posts with images or videos shared on Twitter. For this purpose, they collected supporting documents from two search engines, Google and Baidu, which they then used for veracity evaluation. They considered posts in two languages, English and Chinese. They trained their stance model on English data (FNC-1) using pre-trained multilingual sentence embeddings, and further added cross-platform features in their final neural model.

Nadeem et al. (2019) developed FAKTA, an system for automatic end-to-end fact-checking of claims. It retrieves relevant articles from Wikipedia and selected media sources, which are used for verification. FAKTA uses a stance detection model, trained in a FEVER setting, to predict the stance and to obtain entailed spans. These predictions, combined with linguistic analysis, are used to provide both document- and sentence-level explanations and a factuality score.

Nguyen et al. (2020) proposed the Factual News Graph (FANG) model, which models the social context for fake news detection. In particular, FANG uses the stance of user comments with respect to the target news article, and also temporality, user-user interactions, article-source interactions, and source reliability information.