A Survey on Stance Detection for Mis- and Disinformation Identification

Momchil Hardalov\textsuperscript{1,2} Arnav Arora\textsuperscript{1,3} Preslav Nakov\textsuperscript{1,4} Isabelle Augenstein\textsuperscript{1,3}

\textsuperscript{1}Checkstep Research\textsuperscript{2}Sofia University “St. Kliment Ohridski”, Bulgaria\textsuperscript{3}University of Copenhagen, Denmark\textsuperscript{4}Qatar Computing Research Institute, HBKU, Doha, Qatar

\{momchil, arnav, preslav.nakov, isabelle\}@checkstep.com

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 detecting 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 including 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 (Somassundaran 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., 2022), which mention, though do not exhaustively survey, stance.

However, there is no existing overview of the role that 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 the stances of users or texts towards a claim or a headline (and support fact-checking or studying misinformation) —, or as a component of a system that uses stance as part of its process of judging the veracity of an input claim. Here, we aim to bridge this gap by surveying the research on stance for mis- and disinformation detection, including task formulations, datasets, and methods, from which we draw conclusions and lessons, and we forecast future research trends.
### 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.

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

#### 2 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 personal opinions, but also by other external factors such as cultural norms, roles in the institution of the family, etc. Here, we adopt the general definition of stance detection by 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 3).

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, 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, often written with the intent of ridicule or mockery, and (v) sentiment analysis, which checks the polarity of a piece of text.

#### 3 Stance and Factuality

Here, we offer an overview of the settings for mis- and disinformation detection to which stance detection has been successfully applied. As shown in Figure 1, 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 shows an overview of the key characteristics of the available datasets. We include the source of the data and the target1 towards which the stance is expressed in the provided textual context.

1The 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.

1260
3.1 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 1a.

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, 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, aiming to determine the veracity of a claim based on a set of statements from Wikipedia. Claims can be composite and can contain multiple (contradicting) statements, which requires multi-hop reasoning, and 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 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 the participants were asked to provide new examples to “break” existing models, then to propose “fixes” for the system against such 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 factual evidence for a claim.
FEVER also has a connection to Natural Language Inference, i.e., determining the relationship between two sentences. We 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) studied 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 understanding, and (ii) symbolic, i.e., using the table structure.

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

Fully automated systems can assist in gauging the extent and studying the spread of false information online. This is in contrast to the previously discussed applications of stance detection – as a stand-alone system for detecting mis- and disinformation. Here, we review its potency to serve as a component in an automated pipeline. Figure 1b illustrates the setup, which can also include steps such as modelling the user or profiling the media outlet among others. We discuss in more detail media profiling and misconceptions in Appendix B.

**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 was 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 were among the first to demonstrate the feasibility of this task formulation, the limited size of their study and the focus on assessing the stance of individual posts limited its 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) collected claims and news articles from rumour sites with annotations for stance and veracity by journalists as part of the Emergent project. 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 is the need to summarise, in contrast to FNC-1 (Pomerleau and Rao, 2017), where entire news articles were used.

**Multiple languages**  
While the above research has focused exclusively or primarily on English, 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 Zubiaga et al. (2016b). Bošnjak and Karan (2019) studied stance detection and claim verification of comments for Croatian news articles.
4 Approaches

In this section, we discuss various ways to use stance detection for mis- and disinformation detection, and list the state-of-the-art results in Table 2.

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 3.1. One such line of research is the Fake News Challenge, which used weighted accuracy as an evaluation measure (FNC score), to mitigate the impact of class imbalance. Subsequently, Hanselowski et al. (2018a) criticized the FNC score and F1-micro, and argued in favour of F1-macro (F1) instead. In the competition, most teams used hand-crafted features such as words, word embeddings, and sentiment lexica (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 cues, and found that report and negation are most beneficial, while knowledge and denial are least useful. All these 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 lexica.

Mohtarami et al. (2018) investigated memory networks, aiming to mitigate the impact of irrelevant and noisy information by learning a similarity matrix and a stance filtering component, and taking a step towards explaining the stance of a given claim by extracting meaningful snippets from evidence documents. Like previous work, their model performs poorly on the Agree/Disagree classes, due to the unsupervised way of training the memory networks, 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 Transformers has been explored (Slovikovskaya and Attardi, 2020), significantly improving the performance of previous state-of-the-art approaches. Guderlei and Aßenmacher (2020) showed the most important hyper-parameter to be learning rate, while freezing layers did not help. In particular, using the pre-trained Transformer RoBERTa improved F1 from 18 to 58 for Disagree, and from 50 to 70 for Agree. The success of these models is also seen in cross-lingual settings. For Arabic, Khouja (2020) achieved 76.7 F1 for stance detection on the ANS dataset using mBERT. Similarly, Hardalov et al. (2022) 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 pivotal, 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., 2020b; 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. (2020b); 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, namely, CorefRoBERTa.
Wang et al. (2020) added external knowledge to build a contextualized semantic graph, setting a new SOTA on Snopes. Si et al. (2021) and Ostrowski 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. Atanasova et al. (2022) introduced the task of evidence sufficiency prediction to more reliably predict the NOT ENOUGH INFO class.

Another notable idea is to use pre-trained language models as fact-checkers based on a masked language modelling objective (Lee et al., 2020), or to use the perplexity of the entire claim with respect to the target document (Lee et al., 2021). 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 that the main challenges are (i) confusing semantics at the sentence level, e.g., “Andrea Pirlo is an American professional footballer” vs. “Andrea Pirlo is an Italian professional footballer who plays for an American club.”, (ii) sensitivity to spelling errors, (iii) lack of relation between the article and the entities in the claim, (vi) dependence on syntactic overlaps, e.g., “Terry Crews played on the Los Angeles Chargers.” (NotEnoughInfo) is classified as refuted, given the sentence “In football, Crews played ... for the Los Angeles Rams, San Diego Chargers and Washington Redskins, ...”, (v) embedding-level confusion, e.g., numbers tend to have similar embeddings, “The heart beats at a resting rate close to 22 bpm.” is not classified as refuted based on the evidence sentence “The heart beats at a resting rate close to 72 bpm.”, and similarly for months.

**Threaded Stance** In the setting of 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 further context, 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 3.2, 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) took a step further and 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 to predict the rumour veracity, 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 (Zubiaga et al., 2016b; Derczynski et al., 2017; Gorrell et al., 2019), 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, which goes hand in hand with the threaded structure, 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 robustness. However, setups that combine mis- and disinformation identification with stance detection, outlined in Section 3, vary in their annotation and labelling schemes, which poses many challenges.

Earlier approaches focused on pre-training 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. Recently, Schiller et al. (2021) proposed a benchmark to evaluate the robustness of stance detection 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 models trained on individual datasets. Hardalov et al. (2021) experimented with cross-domain learning from 16 stance detection datasets. They proposed a novel architecture (MoLE) that applies domain adaptation 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. Hardalov et al. (2022) explored PET (Schick and Schütze, 2021) in a cross-lingual setting, combining datasets with different label inventories by modelling the task as a cloze question answering one.

Table 2: 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.\(^2\)

They showed that MDL helps 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.

State of the Art Table 2 shows the state-of-the-art (SOTA) results for each dataset discussed in Section 3 and Table 1. The datasets vary in their task formulation and composition in terms of size, number of classes, class imbalance, topics, evaluation measures, etc. Each of these factors impacts the performance, leading to sizable differences in the final score, as discussed in Section 4, and hence rendering the reported results hard to compare directly across these datasets.

5 Lessons Learned and Future Trends

Dataset Size A major limitation holding back the performance of machine learning for stance detection is the size of the 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 (see Section 4). Several techniques have been proposed to mitigate this, such as sampling strategies (Nie et al., 2019), weighting classes (Veyseh et al., 2017),\textsuperscript{3} crafting artificial examples from auxiliary tasks (Yang et al., 2019; Hardalov et al., 2022), or training on multiple datasets (Schiller et al., 2021; Hardalov et al., 2021, 2022).

Data Mixing A potential way of overcoming limitations in terms of dataset size and focus is to combine multiple datasets. Yet, as we previously discussed (see Section 3), 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. (2021), 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 for 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. (2022) extended that paradigm with sentiment and stance pre-training and evaluated on twelve languages from various domains. 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 improves performance by leveraging better representations learned on a similar language or by acting as a regulariser.

Modelling the Context Modelling the 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 they participate in, 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., 2019; 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 these 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 example are information propagation techniques such as memetic warfare. Hence, it is increasingly important to combine different modalities to understand the full context 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: detecting hateful (Kiela et al., 2020), harmful (Pramanick et al., 2021; Sharma et al., 2022a), and propagandistic memes (Dimitrov et al., 2021a,b); see also a recent survey of harmful memes (Sharma et al., 2022b). This line of research is especially relevant for mis- and disinformation tasks that depend on the wisdom of the crowd in social media as it adds additional information sources (Qazvinian et al., 2011; Zubiaga et al., 2016b; Derczynski et al., 2017; Hossain et al., 2020); see Section 5.
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, but they still do not capture the strength of the expressed stance. As shown in Section 3 (also Appendix 2), 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. Moreover, adding label granularity can further improve the transfer between datasets, as the stance labels already share some semantic similarities, but there can be mismatches in the label definitions (Schiller et al., 2021; Hardalov et al., 2021, 2022).

Explainability  The ability for a model to be able to explain its decisions is getting increasingly important, especially for mis- and disinformation detection, as one could argue that it is a crucial step towards adopting fully automated fact-checking. The FEVER 2.0 task formulation (Thorne et al., 2019) can 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. In a step in this direction, Atanasova et al. (2020a) 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., 2020b), 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 C, 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 3). 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). Nevertheless, these risks can be mitigated by taking into account the credibility of the user or of the information source, which can be done automatically or with the help of human fact-checkers.

6 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.

4http://www.politifact.com/
Acknowledgements

We would like to thank the anonymous reviewers for their useful feedback. Isabelle Augenstein’s research is partially funded by a DFF Sapere Aude research leader grant with grant number 0171-00034B. The work is also part of the Tanbih project, which is developed at the Qatar Computing Research Institute, HBKU, and aims to limit the impact of “fake news,” propaganda, and media bias by making users aware of what they are reading, thus promoting media literacy and critical thinking.

References


1269


Quanzhi Li, Qiong Zhang, and Luo Si. 2019a. event-tAI at SemEval-2019 task 7: Rumor detection on social media by exploiting user content, user credibility and propagation information. In Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval ’19, pages 855–859, Minneapolis, Minnesota, USA.


Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tammy Chakraborty. 2022a. DISARM: Detecting the victims targeted by harmful memes. In Findings of NAACL 2022, Seattle, Washington, USA.


### A Examples of Stance

As outlined in Section 3, there are different formulations in which the task of stance definition is materialised. In Table 3, we present some instances of these as exemplified by different stance detection datasets. The target with respect to which the stance is assessed can vary, e.g., a headline, a comment, a claim, a topic, etc., which in turn can differ in length and form. Moreover, the context where the stance is expressed can vary 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 in (Thorne et al., 2018) and threaded comments in (Gorrell et al., 2019).

In a more detailed view of Table 3, 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 3a shows an example from the News domain, where we have a headline and an entire article body, and the goal is to find how the two are 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 fragments of the input text, which might potentially express different stances. It is possible to simplify the task by extracting a summary of the news article beforehand, and evaluating only the stance of that summary, as shown in Figure 3d. However, obtaining such summaries is not a trivial task: (a) they can be extracted by a human annotator (e.g., a journalist), which is time-consuming and expensive, and can require a priori knowledge about the headline/topic of interest as the article might have more than one highlight or viewpoint, or (b) they can be automatically generated using text summarisation methods, but the result can be noisy.

Stance is often expressed in social media such as Twitter, Facebook, Reddit, etc. We illustrate two such scenarios in Figures 3b and 3e. In contrast to the usually long and well-written news documents, social media posts are mostly short and depend on additional context such as the previous posts in a conversational thread (Figure 3e), or external URLs and implicit topics (Figure 3b). Moreover, these texts also need normalisation, as users tend to use slurs, emojis, and other informal language.

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

Next, in Figure 3c we highlight another interesting setup: claim verification using multiple pieces of evidence. Here, the reasoning is carried in multiple hops over a set of texts. In particular, there might not exist a single passage from a document/post that supports/refutes the claim directly. In that case, a large enough chain of evidence might be needed, which can cover enough contextual knowledge in order to allow the model (or a person) to assess the veracity of the input claim.

Finally, the examples in Figure 3 demonstrate that stance can be used for mis- and disinformation detection in different ways: (i) directly, as in the examples in Figures 3a and 3b, or (ii) as multiple viewpoints, which are later aggregated into a final decision, as in Figure 3c, 3d and 3e.

We thoroughly discussed all of the aforementioned setups in Section 3, including the publicly available datasets that focus on stance in the context of mis- and disinformation identification.

### B Additional Formulations of Stance as a Component for Fact-Checking

Beyond the approaches that we outlined in Section 3.2, stance has also been used for detecting misconceptions and for profiling media sources as part of a fact-checking pipeline. Below, we describe some work that follows these formulations.

**Misconceptions** Hossain et al. (2020) focused on detecting misinformation related to COVID-19, based on known misconceptions listed in Wikipedia. They evaluated the veracity of a tweet depending on whether it agrees, disagrees, or has no stance with respect to a set of misconceptions. A related formulation of the task is detecting previously fact-checked claims (Shaar et al., 2020). This allows to assess the veracity of dubious content by evaluating the stance of a claim regarding already checked stories, known misconceptions, and facts.

**Media Profiling** Stance detection has also been 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.
**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)

**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)

---

The reliability of entire news media sources has been automatically rated based on their stance with respect to 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, e.g., 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, 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. (2019) 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. Finally, Hounsell et al. (2020) proposed to use domain, certificate, and hosting information about the infrastructure of the website. See (Nakov et al., 2021) for a recent survey on media profiling.

There is a well-known connection between factuality and bias. For example, hyper-partisanship is often linked to low trustworthiness (Potthast et al., 2018), e.g., appealing to emotions rather than sticking to the facts, while center media tend to be generally more impartial and also more trustworthy.

**User Profiling** In the case of social media and community fora, it is important to model the trustworthiness of the user. In particular, there has been research on finding opinion manipulation trolls, paid (Mihaylov et al., 2015b) or just perceived (Mihaylov et al., 2015a), sockpuppets (Maity et al., 2017; Kumar et al., 2017), Internet water army (Chen et al., 2013), and seminar users (Darwish et al., 2017).

---

3http://mediabiasfactcheck.com
C Systems and Applications

The systems and applications below use stance detection as part of a pipeline for identifying misinformation, see Section 4 for more details about the methods.

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

Popat et al. (2018) proposed CredEye, a system for automatic credibility assessment of textual claims. The system takes a claim as an input and analyses its credibility by considering relevant articles it retrieved 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. 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 verification: (i) rumour detection, i.e., given a stream of claims, determine whether they are worth verifying or they do contain no rumours, (ii) rumour tracking for finding relevant information about the rumour using social media posts, sentence descriptions, and keywords, (iii) stance classification to collect stances towards the rumour, and (iv) veracity classification to aggregate the information from the tracking component, the collected stances, and optionally other relevant information about the sources, metadata about the users, etc., to predict a truth value for the rumour.

Nadeem et al. (2019) developed FAKTA, a system for automatic end-to-end fact-checking of claims. It retrieves relevant articles from Wikipedia and as well as from selected media sources, which it then uses 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 as an integral component of its model, together with temporality, user–user interactions, article–source interactions, as well a source reliability information.

---

https://gate.d5.mpi-inf.mpg.de/credeye/
http://fcweb.pythonanywhere.com/