Neural Computing and Applications (2023) 35:5113–5144
https://doi.org/10.1007/s00521-023-08285-7

REVIEW

A systematic review of machine learning techniques for stance detection and its applications

Nora Alturayeif[1,2] · Hamzah Luqman[1,3] · Moataz Ahmed[1,4]

Received: 7 October 2022 / Accepted: 6 January 2023 / Published online: 28 January 2023

Abstract

Stance detection is an evolving opinion mining research area motivated by the vast increase in the variety and volume of user-generated content. In this regard, considerable research has been recently carried out in the area of stance detection. In this study, we review the different techniques proposed in the literature for stance detection as well as other applications such as rumor veracity detection. Particularly, we conducted a systematic literature review of empirical research on the machine learning (ML) models for stance detection that were published from January 2015 to October 2022. We analyzed 96 primary studies, which spanned eight categories of ML techniques. In this paper, we categorize the analyzed studies according to a taxonomy of six dimensions: approaches, target dependency, applications, modeling, language, and resources. We further classify and analyze the corresponding techniques from each dimension's perspective and highlight their strengths and weaknesses. The analysis reveals that deep learning models that adopt a mechanism of self-attention have been used more frequently than the other approaches. It is worth noting that emerging ML techniques such as few-shot learning and multitask learning have been used extensively for stance detection. A major conclusion of our analysis is that despite that ML models have shown to be promising in this field, the application of these models in the real world is still limited. Our analysis lists challenges and gaps to be addressed in future research. Furthermore, the taxonomy presented can assist researchers in developing and positioning new techniques for stance detection-related applications.

Keywords Stance detection · Stance classification · Sentiment analysis · Rumor detection · Machine learning · PRISMA

1 Introduction

With the advent of Web 2.0, many online platforms for producing User-Generated Content (UGC) have been established, such as social media, wikis, and debate websites. UGC usually comes in the form of pictures, videos, reviews, or blog posts. Currently, social media platforms are being inherent parts of our daily lives as a media of communication and expressing opinions. Consequently, the amount of available data is rapidly increasing. However, most data are unstructured, where texts represent a substantial part. As the volume of these data increases, the demand for the automatic processing of UGC significantly increases. Advances in machine learning (ML) techniques aid in the extraction of useful information from texts using Natural Language Processing (NLP). This new source of information could be used to measure people's opinions, stances, and attitudes toward products, events, services,

✉ Hamzah Luqman
hluqman@kfupm.edu.sa

Nora Alturayeif
nsalturayeif@iau.edu.sa

Moataz Ahmed
moataz@kfupm.edu.sa

1 Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

2 Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia

3 SDAIA-KFUPM Joint Research Center for Artificial Intelligence, KFUPM, Dhahran, Saudi Arabia

4 Interdisciplinary Research Center of Intelligent Secure Systems (IRC-ISS), KFUPM, Dhahran, Saudi Arabia

controversial news, and politics. These measurements can play a valuable role in decision-making for companies, policymakers, politicians, and even regular people. Furthermore, detecting the stances expressed in a piece of text can be a powerful tool for a range of tasks, such as rumor veracity detection and fake news detection [1, 2].

Stance detection is the task of automatically predicting the writers' stance on a subject of interest (target). It depends on the examination of a written text and sometimes the user's social activity on debate sites (e.g., social media platforms). There are other definitions for stance detection. In the following, we present the definitions of stance detection from different perspectives, and then we will present some related problems.

Before presenting the stance detection definitions, we provide a definition of stance itself from a sociolinguistic perspective. Du Bois [3] defined a stance as *"a public act by a social actor, achieved dialogically through overt communicative means, of simultaneously evaluating objects, positioning subjects, and aligning with other subjects, with respect to any salient dimension of the sociocultural field"*. Kockelman [4] defined it as an expression of the stance taker's attitude and judgment toward a proposition and thereby aligns himself/herself with others. Several definitions of *stance detection* (also known as *stance classification*) can be found in the field of sociolinguistics. The main concern in stance detection is to infer the embedded viewpoint from the authors' text. A study on stance detection is conducted by linking the stance to one or more of the following three factors: linguistic features (tense, lexical aspect, subject, and object), individual identity, and social activity [5]. Stance considerably determines the tone of the writers' message and words that they choose [3].

The term "stance detection" is used in the ML field to refer to a classification problem. The input in this problem is usually in the form of a pair of text and a target, and the output is a category from the set: {Favor, Against, None}. Furthermore, some scholars add to the set the category "Neutral", which implies that the author is neutral toward the target [6]. However, a neutral stance arguably does not exist as people usually position themselves to be against or in favor of a proposition [5]. In addition, there is good agreement in the literature that if the stance of a text toward a target is not in favor of or against it, then the proper stance category would be "None" instead of "Neutral", because no stance information can be obtained from the text. Thus, the "None" category is usually assigned to all cases other than the Favor or Against categories.

In general, stance detection, as observed in the literature, is defined as predicting writers' stance on the target by examining the text they wrote and/or their social activity on social media platforms (connection, preferences, etc.). This definition is schematically illustrated in Fig. 1.

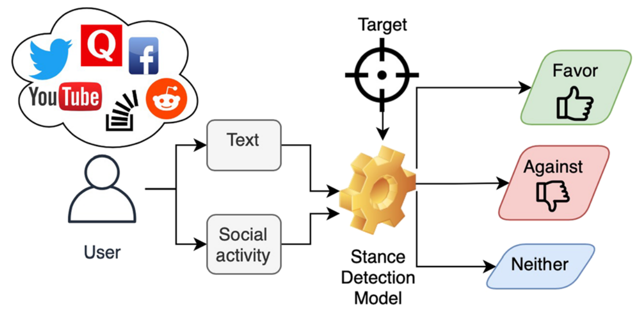

Fig. 1 General representation of the stance detection system

Moreover, stance detection is a problem related to *sentiment analysis* (or *opinion mining*). Sentiment analysis focuses on the sentiment polarity that is explicitly expressed by a text. The main sentiment polarities considered by several scholars are Positive, Negative, and Neutral. By contrast, stance detection aims to classify the stance of a piece of text toward a target (event, entity, idea, claim, topic, etc.) explicitly or implicitly mentioned in the text.

There are two subproblems of sentiment analysis that are more related to stance detection: (1) *Target-dependent sentiment analysis*, and (2) *Aspect-based sentiment analysis*. Both problems are concerned with the identification of the sentiment concerning a specific target (e.g., iPhone vs Galaxy) or different aspects of a target (e.g., screen and battery life of iPhone). It has been noticed in some social media analysis studies that there is a misconception between the definition of stance detection with the generic sentiment analysis as well as the two subproblems (i.e., Aspect-based, Target-dependent) [7]. Thus, we list here the main theoretical differences between them:

- The generic sentiment analysis is concerned with the emotion polarity without a specific target. Meanwhile, in stance detection, a well-defined target must be given to evaluate the position toward this target.
- The stance may not be aligned with the sentiment for a target within a text. That is, a text may have a positive polarity, whereas the stance is against the target, and vice versa. For example, the sentence: *"I am so glad that Trump lost the election"*

  has a positive sentiment, but the stance is against Trump.
- Sentiment analysis studies focus on non-ideological topics (e.g., products and services). Meanwhile, stance detection targets ideological topics (e.g., atheism, feminist movement, and political issues), which are harder to detect.
- In two subproblems of sentiment analysis (i.e., aspect-based, target-dependent), the target in sentiment analysis is usually an entity or an aspect (e.g., reviews about hotels, movies, or products), whereas the target in

stance detection may be an event (e.g., the US presidential election).

Further, stance detection, as a research area in the ML field, is related to other problems besides sentiment analysis. These problems include (i) emotion detection [8], (ii) sarcasm detection [9], (iii) perspective identification [10], (iv) argument mining [11], (v) controversy detection [12], and (vi) biased language detection [13].

Achieving stance detection is challenging due to the fact that determining stance is subjective. In addition, concepts and opinions are formed through a variety of expressions and linguistic compositions, making it more difficult to detect. In social media, stance detection is more demanding due to the nature of social media text [7]. For example, the text is usually short (e.g., a tweet can contain up to 280 characters), informal, containing many abbreviations, and with a nonstandard format due to the users' inconsistent use of grammar. Furthermore, social media discussions are more scattered and lack contextual information [14].

An increasing number of research papers and applications are published by multiple communities on the stance detection problem. With this large number of studies, there is a need to have a framework to classify the available approaches in the literature, since they use various techniques and rely on different underlying models. This is crucial to enable researchers and practitioners to understand the contexts of the different approaches and their suitability for different circumstances. Furthermore, there are still open gaps and promising future trends to be explored toward more robust stance detection models. This study aimed to propose a framework for classifying different approaches, evaluating the current state of affairs, and identifying open gaps.

In this study, we present a systematic literature review (SLR) that focuses on stance detection. To the best of our knowledge, there is no SLR in this area, which motivates our work in the current study. Our research investigates the ML techniques used in the literature for stance detection by addressing five research questions following a well-defined methodology. The contributions of this SLR lay on:

- Covering the most recent studies (2015–2022) and a significant number of papers resulting from an established literature review protocol.
- Proposing a taxonomy to classify the literature on the stance detection domain, as well as a taxonomy of different techniques used for stance modeling.
- Classifying 96 selected studies according to the proposed taxonomy.
- Summarizing the current state-of-the-art stance models with a focus on ML techniques.
- Introducing open gaps to be explored for future research toward more robust approaches for stance detection.

The rest of this article is organized as follows. Section 2 presents previous works regarding literature reviews related to stance detection. Section 3 presents the methodology used for performing the present SLR, starting with our research questions. Section 4 presents and discusses our results from this SLR by addressing the research questions. Finally, Sect. 5 concludes this survey.

2 Related reviews

Survey studies can be broadly divided into two categories, namely, traditional literature reviews and SLRs [15]. Traditional literature reviews mainly cover the research trends, whereas the SLRs aim to answer various research questions. In the field of social computing, stance detection is a comparatively recent computational problem. Although the fact that there are multiple survey studies in this newly established area [2, 7, 16–18], there is no existing SLR in this domain, which motivates our work in this survey.

Furthermore, some of these survey studies targeted only one aspect of stance detection. Hardalov et al. [16] surveyed the applications of stances for misinformation and disinformation detection. Alkhalifa and Zubiaga [17] investigated the existing directions in capturing stance dynamics in social media. They reviewed the relevant literature on the temporal dynamics of social media and discussed their impact on the development of stance detection models. Wang et al. [18] surveyed the opinion mining methods in general, with a particular focus on customers' stances toward products. Their study emphasized the methods for extracting textual features of social media posts only, where they examined numerous techniques for extracting aspects from posts commenting about products.

Relatively comparative surveys in stance detection were published in 2019 and 2020 by Küçük and Can [2] and Aldayel and Magdy [7], respectively. Küçük and Can [2] discussed the NLP techniques used with stance detection. Their survey includes a useful explanation for the intersections and distinctions between stance and related tasks, such as emotion recognition, sarcasm, and argument mining. Aldayel and Magdy [7] surveyed studies on stance detection targeting the social media domain, starting by providing a broad overview of the stance detection task, including the definition, theoretical comparison between stance and sentiment, feature modeling, and different types of stance targets. Then, they presented a breakdown of the most recent approaches to stance modeling in social media.

The related reviews presented above are limited by study selection bias as they did not seem to follow a systematic selection methodology. Moreover, those studies are not comprehensive as they seemed overly restrictive in terms of the approaches and applications considered. These shortcomings motivated us to conduct this SLR that comprehensively explores and analyzes relevant prominent studies from different domains and applications. This SLR also outlines the present literature gaps and suggests possible research directions to improve the current state of the affairs. In addition, related reviews did not deeply discuss the emerging techniques of machine learning (e.g., inductive transfer learning and low-shot learning) as presented in this survey.

3 Methodology

In this study, we compile, categorize, and present a comprehensive and up-to-date survey of stance detection models and applications. To enforce sound inclusion eligibility criteria, we followed the SLR procedure proposed by Kitchenham [19]. The main advantage of this procedure over others is that it was designed primarily for computer science surveys, which helps in adapting it well to the stance detection topic. In addition, following this well-defined protocol makes the study reproducible and reduces the possibility of bias in the results of the literature.

During the planning stage of this SLR, we developed a review protocol that is broken down into five phases: research question definition, search strategy design, study selection, quality assessment, and data extraction. Details of the review protocol will be presented in the following subsections.

3.1 Research questions

The goal of our study is to answer the following research questions:

$RQ1$: What is the current state of the stance detection research?

$RQ2$: What taxonomy could be used to represent the stance detection applications?

$RQ3$: What is the focus of the stance detection research? Particularly, what are the platforms and domains for which stance detection models were proposed? How is stance modeled in the selected studies?

$RQ4$: What are the major developments in the stance detection research? Particularly, what are the ML techniques used and how can they be classified?

$RQ5$: What are the research gaps observed in the literature?

3.2 Search strategy

Preliminary searches were performed to determine the number of possibly relevant studies in the stance detection area. When we applied the query by searching full texts, an unfeasible volume of irrelevant papers was returned (hundreds of thousands) as the searched phrases are common in other fields (e.g., sociolinguistics). Therefore, we have decided to conduct our search based on title, abstract, and keywords. In addition, we used alternative terms and synonyms for the topics we were looking for throughout our preliminary searches. As a result, the following query string was used for identifying primary studies:

"stance detection" OR "stance prediction" OR "stance identification" OR "stance classification" OR "stance recognition".

We restricted the search to the period from January 2015 to October 2022. The period constraint was set due to the significant increase in the number of studies that targeted stance detection compared with the studies published before 2015. Furthermore, most of the techniques proposed prior to 2015 relied primarily on statistical modeling rather than machine learning for stance detection.

The following electronic libraries were selected as sources for our study: ACM Digital Library[1], Scopus[2], Springer[3], Web of Science[4], and IEEE-Xplore[5]. These libraries were selected because they host the major journals and conference proceedings related to social computing and ML. To complement these libraries, we also searched Google Scholar[6]. Consequently, a total of six libraries were examined in this SLR.

In addition, we conducted backward snowballing by scanning the references in the relevant papers. We identified ten extra papers, and three of them were found to be relevant and passed the quality assessment (presented in Sect. 3.3). These papers have been included in the final number of selected papers.

The applied search strategy was based on preferred reporting items for systematic review and meta-analysis (PRISMA) statements [20], which is summarized in Fig. 2. The search results were managed and stored using the Mendeley software package (https://www.mendeley.com/). According to the search procedure, we identified 96 primary studies out of 654 studies that resulted from the first search phase. Figure 2 presents the detailed search procedure as well as the number of papers found at each phase.

[1] https://dl.acm.org/.

[2] https://www.scopus.com/.

[3] https://link.springer.com/.

[4] https://mjl.clarivate.com/.

[5] https://ieeexplore.ieee.org/.

[6] https://scholar.google.com/

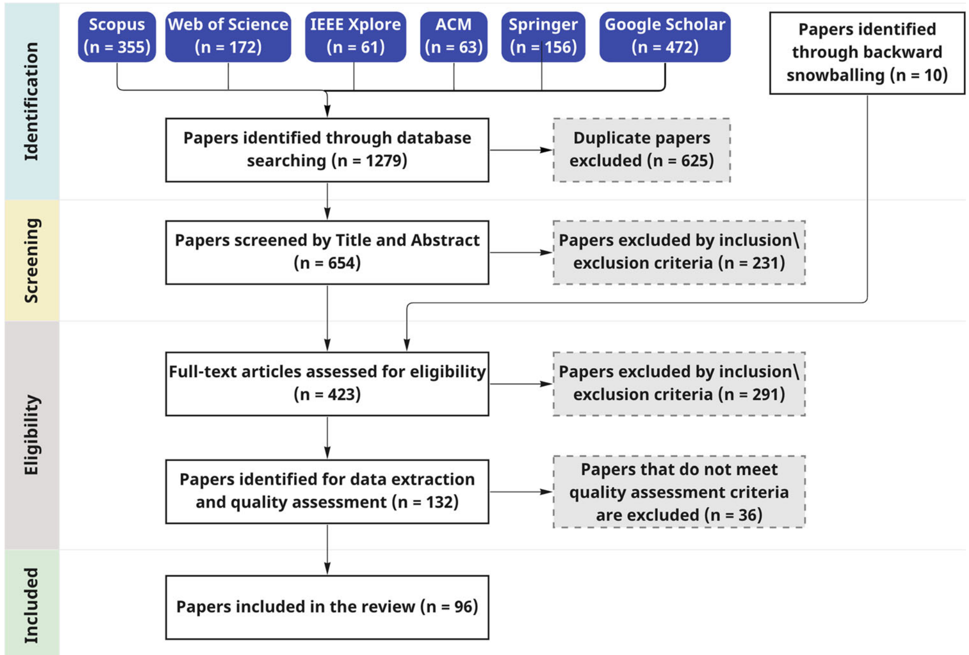

Fig. 2 PRISMA flow diagram for the search strategy; where *n* is the number of papers

3.3 Study selection

The first search phase resulted in 654 candidate papers (see the identification phase in Fig. 2). These papers obtained from the identification phase were evaluated by applying the inclusion and exclusion criteria to identify the most relevant papers for our SLR. Papers that met all inclusion criteria were included, whereas those that met any exclusion criterion were excluded. The following inclusion and exclusion criteria were developed and refined through a pilot selection. We selected papers by looking at their titles, abstracts, and full texts.

- Inclusion criteria
  1. Empirical studies using the ML techniques for stance detection, whether as a main task or as an auxiliary task for other applications.
  2. Papers that study the detection of stance based on the following forms of data: text and social media networks.
  3. Papers written in English.
  4. Peer-reviewed papers.
  5. In the case of multiple publications of the same study, only the most recent and comprehensive version was included.
  6. For notebook papers of the annual SemEval workshop and other competitions related to stance detection, only the top two papers (based on the reported results referenced in the official overview papers of the workshops) were included.
- Exclusion criteria
  1. Papers that do not satisfy any of the specified inclusion criteria.
  2. Survey or review papers without any findings.
  3. Extended abstracts, posters, books, patents, tutorials, and short papers (as categorized by conferences).
  4. Inaccessible papers.
  5. Studies focusing on building a resource for stance detection, such as datasets, lexicons, annotation framework, or solutions for addressing imbalanced data.

The use of these selection criteria resulted in the identification of 132 studies. The final selected studies were obtained using the quality assessment criteria, which we formed for evaluating the relevance and strength of the main studies. The quality assessment criteria are listed in Table 1. The questions are ranked as follows: “Yes” = 1, “Partly” = 0.5, and “No” = 0. After summing the values assigned to each question, the total score is calculated. A study could have a maximum score of 8 and a minimum score of 0. We considered only the relevant studies with a quality score greater than 4 (i.e., 50% of the maximum score), which were eventually used for data extraction. Accordingly, we further dropped 36 relevant papers with a quality score of 4 or less. Consequently, 96 studies were finally identified for the data extraction process.

3.4 Data extraction and data synthesis

Relevant data were extracted from each of the selected papers in order to fulfill RQs 1–5. In addition, we collected

Table 1 Quality assessment questions

Q#	Quality questions
Q1	Does the paper have a well-defined methodology?
Q2	Is the information about the dataset size and data source identified?
Q3	Are the pre-processing techniques clearly described and justified?
Q4	Are the ML techniques sufficiently defined?
Q5	Are the performance measures fully defined and reported?
Q6	Is there a comparison with other approaches?
Q7	Does the study add/contribute to academia?
Q8	Does the study have sufficient number of the average citations per year?

the metadata information on each paper for further statistical investigation. The metadata included the title, publication year, authors, type of publication, venue, and the number of citations. The extracted data were organized using Excel spreadsheets.

The primary goal of data synthesis is to collect and combine facts and statistics from the selected studies to answer RQs 1–5 and build a response. Grouping studies with similar and comparable outcomes helped obtain conclusive answers to RQs by presenting research evidence. We examined both quantitative and qualitative data, such as prediction accuracy, approach category, feature extraction technique, ML method, language, domain, and dataset. To synthesize data from the primary studies and address RQs 1–5, various techniques were used, including visualization techniques (e.g., treemap and word cloud). Tables were also used to summarize and present the findings.

4 Results and discussion

In this section, we present and discuss the results of our literature analysis. In each of the following five Sects. (4.1, 4.2, 4.3, 4.4, 4.5), we present and discuss our findings inline with RQs 1–5.

4.1 The current state of research on stance detection (RQ1)

The objective of this section is to answer RQ1, which is related to showing the current research state on stance detection. Therefore, we start by presenting the population of the published literature on stance detection and the leading publication venues. In addition, we survey the competitions (shared tasks) related to stance detection in Sect. 4.1.2. Furthermore, we present the datasets and resources used in the current stance detection models in Sect. 4.1.3.

4.1.1 Description of primary studies

Stance detection (also known as stance classification, stance identification, and stance prediction) is a considerably recent computational problem in the area of social computing. One of the observations during our literature review is the significant growth in the number of studies on the stance detection topic in recent years. Figure 3 presents the number of stance detection publications and the publications from 2015 to 2022 after applying the SLR protocol (presented in Sect. 3). It can be observed from the figure that there is a noticeable growth in the number of publications from 2016, which is attributable to the publication of the SemEval-2016 competition that presented the first benchmarked dataset for stance detection based on social media contents [21]. This dataset opened up opportunities to develop models for stance representations on social media.

We selected 96 of 654 identified papers that used ML techniques for stance detection (based on the SLR protocol presented in Sect. 3). About 21% of these papers were issued in journals, and the rest were published in conference proceedings. Table 2 presents the publication venues and distribution of the papers per venue. As presented in Table 2, the top two publication venues are EMNLP and ACL conferences, with around 23% of the selected papers (14% and 9%, respectively). Both conferences are prestigious in the computational linguistics field, where substantial advances in NLP are likely to be published.

4.1.2 Stance detection competitions

Besides the SemEval-2016 competition mentioned earlier, there are six competitions have been held for stance detection. All these competitions contributed to the advancement of stance detection research by offering annotated datasets of different languages, annotation guidelines, evaluation metrics, and an overview of the participating teams. The details of these competitions are presented next in chronological order. Furthermore, the

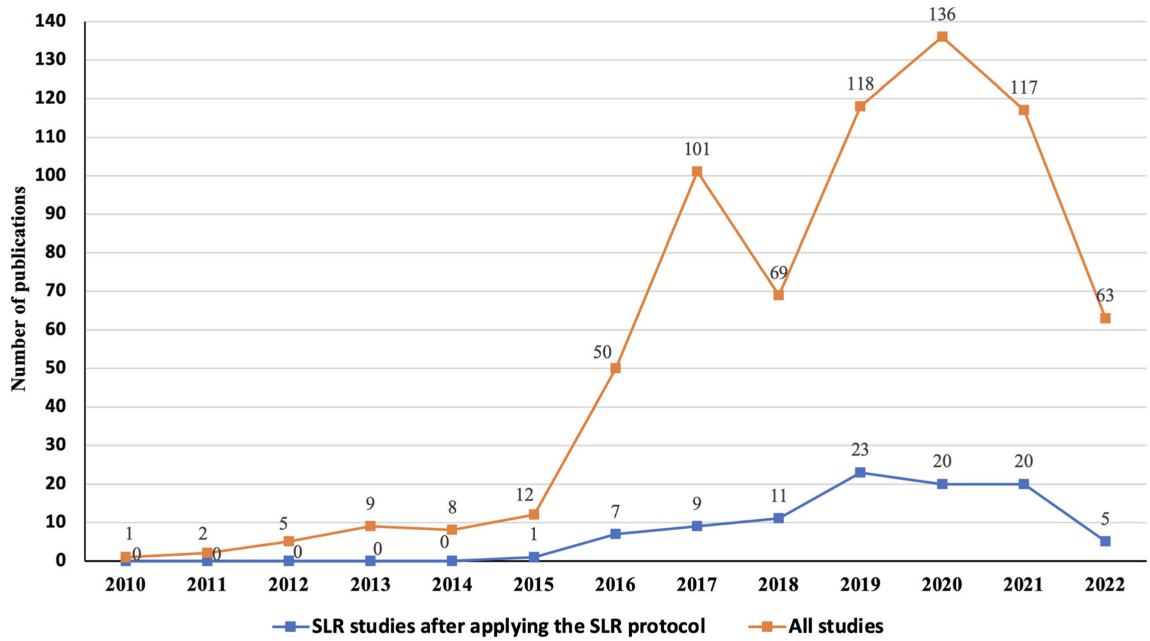

Fig. 3 Number of stance detection studies between 2010 and 2022

Table 2 Publication venues and the distribution of selected studies

Publication Venue	Type	# studies	Percent
Empirical Methods in Natural Language Processing (EMNLP)	Conference	13	13.54
Association for Computational Linguistics (ACL)	Conference	9	9.38
International Workshop on Semantic Evaluation (SemEval)	Conference	6	6.25
International Conference on Computational Linguistics (COLING)	Conference	4	4.17
IEEE ACCESS	Journal	3	3.13
World Wide Web Conference (WWW)	Conference	3	3.13
ACM Transactions on Information Systems (TOIS)	Journal	2	2.08
IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining	Conference	2	2.08
Information Processing & Management	Journal	2	2.08
International AAAI Conference on Web and Social Media (ICWSM)	Conference	2	2.08
International Conference on Artificial Neural Networks (ICANN)	Conference	2	2.08
International Conference on Data Mining (ICDMW)	Conference	2	2.08
International Conference on Natural Language Processing and Chinese Computing	Conference	2	2.08
International Joint Conference on Neural Networks (IJCNN)	Conference	2	2.08
Other journals	Journal	15	15.63
Other conferences	Conference	27	28.13
Total		96	100

information on the datasets used in these competitions is presented in Table 3.

1. SemEval-2016 Task 6 (SE16-T6): This is the first shared task on stance detection that was organized as a part of the International Workshop on Semantic Evaluation [21]. The competition comprised two subtasks: Tasks A and B. Task A is a supervised stance detection in English Tweets where the participants are provided with 70% of annotated training data. Task B is a weakly supervised stance detection where the participants are given only a large unlabeled dataset along with a smaller test dataset for a new target. Notably, in this competition, the worst performing systems are based on deep learning methods [21]. It

Table 3 Publicly available stance detection datasets

Dataset Name	Language	Target depen.	Domain	Targets	Annotation	Dataset Size
Emergent [27]	English	TI	Claims from different sites	Several topics	Favor, against, observe	300 claims and 2,595 articles
SemEval-2016 Task 6 [21]		TS	Tweets	Atheism, Climate change, Feminist movement, Hillary Clinton, Abortion	Favor, against, none	4163 tweets
Multi-Target SD [28]		MrT	Tweets	Donald Trump, Ted Cruz, Hillary Clinton, Bernie Sanders	Favor, against, none	4455 tweets
IBM Debater [29]		TI	Claims and evidence from Wikipedia	55 topics	Pros, cons	2394 claims
RumourEval-17 [24]		TI	Tweets	Rumors about ten events	Support, deny, query, comment	5568 tweets
FNC-1 [30]		TI	News headlines	Several topics	Agree, disagree, discuss, unrelated	2587 news headlines
UKP or AM [11]		TS	Posts from debate websites	Several topics	Favor, against, none	25,492 comments
Perspectrum [31]		TI	Posts from debate websites	Several topics	Support, opposing	11,876 pairs (perspective, claim)
Args.me [32]		TI	Posts from debate websites	Several topics	Pros, cons*	387,606 arguments
RumourEval-19 [25]		TI	Tweets, Reddit posts	Natural disasters	Support, deny, query, comment	8574 posts
VAST [33]		CT	Posts from The New York Times	Several topics	Pros, cons, neutral	23,525 comments
WT-WT [34]		TS	Tweets	Health insurance companies	Support, refute, comment	51,284 tweets
TW-BREXIT [35]		TS	Tweets	BREXIT referendum	Leave, remain, none	1800 triplets of tweets
Procon20 [36]		TS	Posts from procon.org	419 controversial issues.	Pros, cons	6094 pairs (question, opinion)
Grimminger et al. [6]		TS	Tweets	Donald Trump, Joe Biden, Kanye West	Favor, against, none, hateful, non-hateful	3000 tweets
Baly et al. [37]	Arabic	TI	Posts from Verify and Reuters	War in Syria and related political issues	Agree, disagree, discuss, unrelated	422 claims and 3,042 articles
Arabic News Stance [38]		TI	News headlines	Several topics	Agree, disagree, other	3786 pairs (claim, evidence)
ConRef-STANCE-ita [39]	Italian	TS	Tweets	The reform of the Italian Constitution	Favor, against, none	963 triplets (tweet, retweet, reply)
SardiStance [26]		TS	Tweets	Sardines movement	Favor, against, none	3242 tweets
NLPCC-2016 Task 4 [22]	Chinese	TS	Weibo posts	Several topics	Favor, against, none	3250 posts
Hercig et al. [40]	Czech	TS	News comments	Miloš Zeman, Smoking ban in restaurants	Favor, against, none	5423 comments
KÜÇÜK et al. [41]	Turkish	TS	Tweets	Football clubs	Favor, against	1065 tweets

Table 3 (continued)

Dataset Name	Language	Target depen.	Domain	Targets	Annotation	Dataset Size
Pheme [42]	Multi (English, French, German)	TI	Tweets	Rumors about nine events	Support, deny, query, comment	4842 tweets
X-stance [43]	Multi (French, German, Italian)	CT	Posts from Smartvote website	150 political issues	Favor, against *	German: 40,200, French: 14,129, Italy: 1,173
IberEval 2017 [23]	Multi (Catalan, Spanish)	TS	Tweets	Catalan Independence	Favor, against, none	5400 tweets (for each language)
Zotova et al. [44]		TS	Tweets	Catalan Independence	Favor, against, none *	Spanish: 10K, Catalan: 10K

The * in the annotation column means that the dataset is annotated automatically

has been hypothesized that due to the irregular syntax of social media text and the small size of training data, traditional deep learning methods cannot model tweet text well.

2. NLPCC-2016 Task 4: For Chinese microblogs, a stance detection competition was held with two subtasks similar to SE16-T6 [22].
3. IberEval-2017: A shared task conducted for stance and gender detection in Spanish and Catalan tweets [23].
4. SemEval-2017 Task 8 (RumourEval-2017): A shared task aimed at identifying rumors and the stance of Twitter users through their textual replies [24].
5. SemEval-2019 Task 7 (RumourEval-2019): A shared task that comprised two tasks: rumor verification and rumor stance prediction on Twitter and Reddit posts [25].
6. EVALITA-2020 (SardiStance): SardiStance, held during the EVALITA-2020 conference, was the first shared task for stance detection in the Italian language [26]. This competition also comprised two subtasks: Tasks A and B. Task A is related to textual stance detection, and Task B is based on contextual stance detection that uses additional information from the user's social network and tweets, as well as information about the user profile.

4.1.3 Resources

In this SLR, we also reviewed the resources that were employed across all selected studies for stance detection. These resources involve datasets, lexicons, and knowledge graphs. Although stance classification is a recent research area, extensive effort is dedicated to creating and annotating datasets for this task. The annotated datasets have been used to train both supervised and weakly supervised models. In addition, they have been used for validating unsupervised models. In the surveyed literature, we encountered many public stance detection datasets of different text types (news headlines, news comments, tweets, and posts in online forums). The datasets targeted ten languages: Arabic, Catalan, Chinese, Czech, English, French, German, Italian, Spanish, and Turkish.

Table 3 presents the details of the surveyed datasets in terms of language, target dependency (TS: target-specific, MrT: multi-related targets, CT: cross-target, and TI: target-independent), domain, targets, annotation classes, and dataset size. We only included the publicly available datasets that are listed in chronological order in Table 3. The table lists 26 public datasets, 6 of them are shared-task datasets: NLPCC-2016 Task 4, SE16-T6, RumourEval-17, RumourEval-19, SardiStance, and IberEval-2017. It is worthwhile noting that 55 of the 96 reviewed studies considered shared-task datasets. SE16-T6 is the most dominant one and was used by 38 studies.

Aside from the datasets, different lexicons (e.g., VADER [45]) were used by 13 studies (out of 96). These lexicons were used as extra features to train ML models. In the following, we list the top eight lexicons along with the studies that used them (note that some studies used more than one lexicon).

1. NRC (also known as EmoLex)[7] [46]: an emotion lexicon used in [36, 47–51].
2. Hu and Liu[8] [52]: an opinion lexicon used in [35, 48, 50, 53–55].
3. MPQA[9] [56]: a subjectivity lexicon used in [48, 50, 51, 53, 57].
4. LIWC (Linguistic Inquiry and Word Count)[10] [58]: an emotion lexicon used in [35, 47, 55, 59].

[7] https://www.saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

[8] https://www.cs.uic.edu/%7eliub/FBS/sentiment-analysis.html.

[9] https://mpqa.cs.pitt.edu/lexicons.

[10] http://liwc.wpengine.com.

5. DAL [60]: an emotion lexicon used in [35, 51, 55].
6. AFINN (Affective Norms for English Words)[11] [61]: a sentiment lexicon used in [35, 51, 55].
7. VADER (Valence Aware Dictionary and sEntiment Reasoner)[12] [45]: a lexicon and rule-based sentiment analysis tool used in [36, 59].
8. SenticNet[13] [62]: a semantic lexicon used in [47, 49].

In addition to the aforementioned lexicons, only one study created a new lexicon as part of their work. The authors of [54] constructed a stance lexicon[14] to guide the attention mechanism in their stance detection model. Specifically, they built a stance lexicon for each target in the SE16-T6 dataset as well as 1000 additional tweets that have been collected using specific hashtags for each target.

External knowledge graphs are another resource used for stance detection. Two studies (out of 96) used this resource [63, 64]. Both studies adopted the ConceptNet knowledge graph [65], which comprises millions of relation triples (head concept, relation, and tail concept). ConceptNet was used to construct relational subgraphs for building a commonsense knowledge-enhanced module to be used by low-shot techniques for stance detection.

4.2 Stance detection taxonomy (RQ2)

The second research question that we are trying to answer in this survey is "What taxonomy could be used to represent the stance detection applications?" Aiming to answer this question, we propose a taxonomy of research work in stance detection which is shown in Fig. 4. As depicted in the figure, the reviewed studies can be classified in six dimensions: *ML approaches*, *target dependency*, *applications*, *modeling* (stance representation), *language*, and *resources*. The number of studies belonging to each dimension is presented in Fig. 4. It should be noted that each study can fit into all the different dimensions. In addition, there is no overlap between branches (i.e., categories) within a dimension. Meaning that we can describe each study using a category from each of the six dimensions. In the following, we describe each dimension:

ML approaches Existing approaches for stance detection can be broadly categorized into two based on feature extraction and learning: non-machine learning (or feature-based) and machine learning (or data-driven) techniques. The non-machine learning approaches involve techniques that depend on hand-crafted features to represent the stance (e.g., arguing lexicon and social activity). These techniques have been employed by some studies in the literature; however, we excluded them during the inclusion and exclusion stage of the SLR protocol. Meanwhile, data-driven techniques use machine learning or deep learning algorithms to train a classifier in a supervised, weakly supervised, or unsupervised manner. Some studies combine both approaches for the stance detection problem [35, 55, 66]. More details on the different ML techniques for stance detection are presented in Sect. 4.4.

Target Dependency Target dependency in stance detection studies can be categorized into four: target-specific (or specific target), multi-related targets, cross-target, and target-independent (as shown in Fig. 4). In *target-specific* studies, the text or the user is the main input to identify the stance toward specific and predefined targets, such as Donald Trump in the US election and the BREXIT referendum. Few studies considered *multi-related targets* by applying one stance detection model to multiple related targets. In these studies, it was assumed that when people express their stance on one target, they indicate their stance toward the other related targets (e.g., Trump versus Biden).

In both target-specific and multi-related targets studies, the task's boundary is defined by the target on which the stance is taken, and training data for every target are usually given for prediction on the same target. However, in *cross-target* studies, researchers investigate the possibility of generalizing classifiers across targets. The objective of cross-target systems is to propose models that can transfer learned knowledge between targets (from a source target to a destination target), for instance, training a classifier on "Donald Trump" and predicting on "Joe Biden". For the *target-independent* studies, in which the target of the stance is not an explicit entity. In fact, the target in these studies is a claim in a piece of news. Target-independent models aim to detect the stance in the comments about some news (confirming the news or denying its validity), or to predict whether a given pair of arguments argue for the same stance (i.e., same side stance classification). Table 4 lists the surveyed studies categorized by target dependency and publication year; most studies targeted a specific topic (target-specific). Meanwhile, there are few studies on multi-related targets due to the challenges associated with this task and the lack of annotated datasets.

Applications The applications of stance detection (other than identifying the stance of a user toward some target) can be categorized into three: rumor veracity detection, fake news detection, and diachronic evolution analysis. In the *rumor veracity* task, a stance detection model is used to determine the veracity of a currently circulating story or information that is yet to be verified at the time of spreading [113]. In more formal terms, given a pair of textual rumors and responses, stance detection refers to the classification of the text's position toward the rumor into a

[11] https://github.com/fnielsen/afinn.

[12] https://github.com/cjhutto/vaderSentiment.

[13] https://www.sentic.net

[14] https://github.com/chuchun8/EMNLP19-Stance.

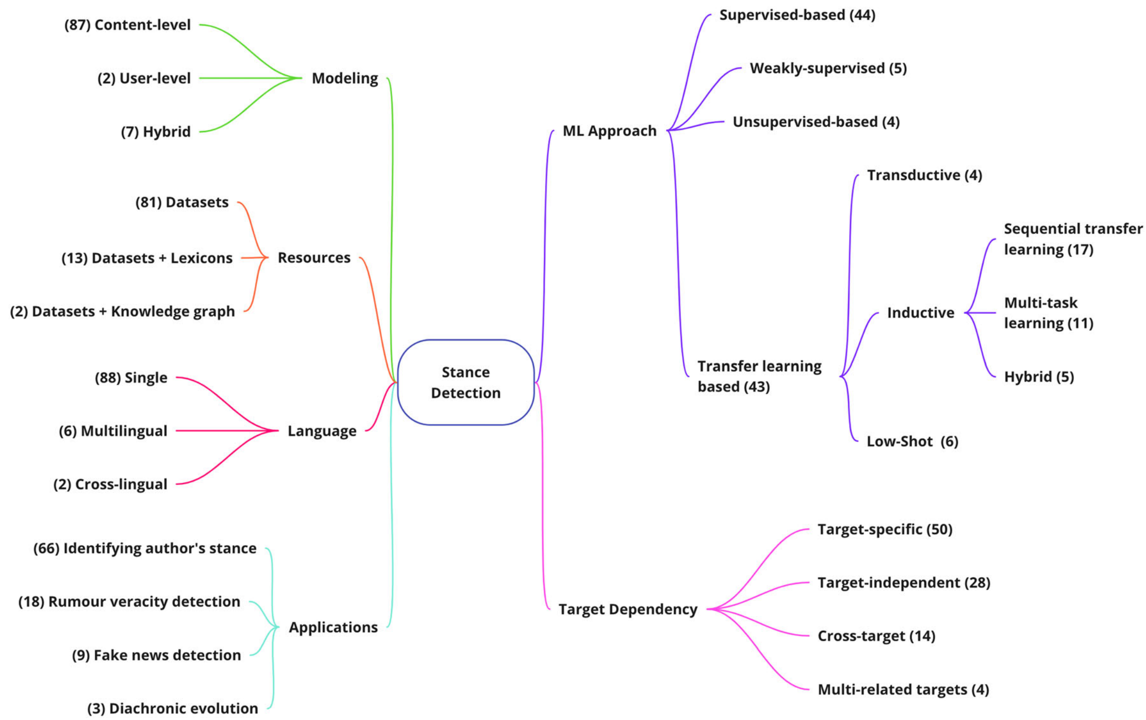

Fig. 4 Proposed taxonomy of the Stance Detection problem with the number of surveyed studies in each subcategory

Table 4 Selected studies categorized by target dependency and publication year

Target dependency	2015–2016	2017–2018	2019–2020	2021–2022
Target-specific	[48, 59, 67–71]	[50, 57, 72–79]	[35, 36, 51, 54, 55, 66, 80–99]	[100–106]
Target-independent		[42, 107–113]	[38, 47, 53, 114–124]	[125–130]
Cross-target	[131]	[132]	[33, 43, 49]	[63, 64, 133–139]
Multi-related targets		[140]	[141, 142]	[143]

label from the set {Support, Deny, Query, Comment}. This configuration has been widely examined in the context of social media microblogs [144]. *Fake news* detection is a similar field in which the veracity of circulating information does not need to be confirmed at the time of dissemination, as the fake news is intentionally written to mislead consumers. Thus, the task is to detect news that is always fake and contains specific types of misinformation. A well-known example of this task is to determine the relationship between a headline and the content of an article (probably from another news source). The possible classes for this task are Agree, Disagree, Discuss, and Unrelated. However, the challenges of recognizing fake news and rumors are essentially the same; usually, auxiliary information, such as user credibility on social media, is required to make a decision.

Moreover, the analysis of *diachronic evolution* is a recent research area in stance detection, in which the researchers explore the stance toward a specific target at the user level by aggregating data over time, considering different time-window sizes [101]. This task is usually defined as a three-way classification where each post is assigned to a stance in favor, against, or neutral. The goal of studying diachronic evolution is to understand the temporal variations in the real world and their impact on public opinion. Developing models for this task requires large datasets collected over different periods of time. Table 5 provides some examples of input formulation with the corresponding target and stance polarity in different stance detection applications.

Modeling Modeling the features of stance on social media can be classified into three: content-level, user-level, and hybrid. The *content-level* modeling is modeled by the linguistic features (e.g., topic modeling, N-gram, and word embeddings) and sentiment information. *User-level* features include the users' interactions, preferences,

Table 5 Examples of input formulation with the corresponding target and stance in different stance detection applications

Application	Input formulation	Target	Stance	Ref.
Identifying author's stance	Tweet (e.g., "The woman has a voice. Who speaks for the baby? I'm just asking")	Legalization of abortion	Against	[50]
Diachronic evolution analysis	Tweets from different time-window (six-year time period)	Gender equality	Favor, against, or none	[101]
Rumor veracity detection	Tree-structured thread discussing the veracity of a source tweet introducing a rumor	NA	Support, deny, comment, or query	[128]
Fake news detection	News headlines and a set of articles	NA	Agree, disagree, discuss, or unrelated	[126]

connections, and timelines on their social platforms. *Hybrid* models learn representation from both content and user features. The details of the features used for stance modeling are presented in Sect. 4.3.3.

Language The literature on stance detection can also be categorized based on the targeted language: single language, multilingual, and cross-lingual. However, most studies on stance detection target a *single language*. English is the main language targeted by most stance detection studies; only a handful of stance detection studies considered languages other than English. In *multilingual* studies, researchers create one model for different languages using datasets for each language. For stance detection in a *cross-lingual* setting, the domain adaptation approach is generally considered when there are sufficient labeled data in one language, and the aim is to learn representations from this language that are useful for another language with few learning data.

Resources Different types of resources have been used in the literature for stance detection. The three main forms of these resources are *datasets (labeled or unlabeled)*, *lexicons* (e.g., VADER for sentiment polarity [45]), and *knowledge graphs* (e.g., ConceptNet [65]) used in [63, 64]. The details of these resources are presented in Sect. 4.1.3.

4.3 Context of stance detection studies (RQ3)

In this section, we aim to answer RQ3 by presenting the focus of the stance detection research in terms of the platforms and domains for which stance detection models are proposed and how the stance is modeled in the selected studies. Section 4.2 presented the different aspects that were adopted in the selected studies. In this section, we show how the tasks were implemented for three aspects: platforms, domain areas, and stance modeling.

4.3.1 Platforms

Several platforms have been used in the literature as data sources for model training and evaluation. The main platforms that have been used in the literature for stance detection are social media, news websites, and debate websites. Figure 5 presents the percentage of studies per platform; notably, 9% of the selected studies adopted multiple types of platforms. From the figure, most selected studies adopted social media platforms as their context for building models. Twitter is the most used dataset resource; it was used by 64 of the 96 selected studies. Meanwhile, only two studies used Weibo [75, 93], and one study considered Facebook [68]. These findings highlight the significance and popularity of social media for research and development in this field. The high dependency on Twitter can be attributed to the accessibility and ethical considerations in data extraction using Twitter APIs compared with other social media platforms (e.g., Facebook) that pose more challenges in data extraction.

Moreover, ten studies considered debate websites to collect data and evaluate their models. For example, *www.procon.org* is used in [36, 88] to collect a set of controversial issues and their related pros and cons posts. This resulted in long documents with numerous words per document (the average number is 166 words) in contrast to data collected from social media that use samples with fewer words (Twitter uses a maximum of 280 characters per sample). The websites *www.Idebate.com* and *www.debatewise.org* were also considered in [116] to have a set of controversial claims and users' perspectives in order to infer these perspectives in terms of supporting or opposing the claim. Although these debate websites are being used as a resource to encourage critical thinking and present information in a nonpartisan format, the topics covered are limited, and the training data would not extend to general topics, such as those discussed on social media platforms.

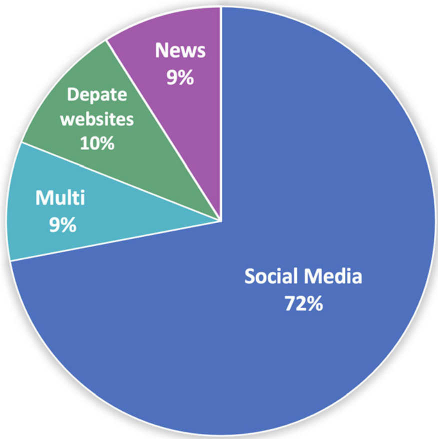

Fig. 5 Distribution of the data source platforms used by the selected studies

The news domain has been considered by several studies. This type of platform is considered mostly by *target-independent* studies. The models in these studies were built for fake news detection or rumor veracity detection tasks. In fake news detection studies, the models are depending on the news headlines and body texts to evaluate the stance of the body text toward a specific target. Several polarities have been targeted in these studies such as agrees, disagrees, discusses, and unrelated. On the other hand, the typical input for rumor veracity models is a stream of social media posts that report circulating news or story. The goal of these models is to classify each post as a rumor or not rumor.

4.3.2 Domain area

Most selected studies focused on one or more controversial topics. Figure 6 presents a treemap that shows the main domains that have been targeted by stance detection studies, where the sizes of the rectangles represent the number of studies. The main domains, as shown in Fig. 6, are political issues (e.g., the US election), social issues (e.g., feminist movement), health (e.g., COVID-19 vaccine), and science (e.g., climate change) issues. However, some studies do not target-specific topics, where their proposed models are designed to detect the veracity of rumors/fake news in general or to assess the position of a claim toward any topic.

The political or government domain is the dominant topic area targeted by most stance detection approaches. These approaches are applied to different political events or actors, such as Hilary Clinton (all studies that considered the SE16-T6 dataset), the Turkish election [94, 95], the war in Syria [75, 83, 87, 93], Catalan independence [55, 74], the US presidential candidates [103, 140–142], gun control and rights [80, 96], and the BREXIT referendum [35, 90]. In terms of the social domain, all studies that considered the SE16-T6 dataset evaluated their models on two social topics: atheism and the feminist movement. In addition, some other studies focused on gay rights [68, 125] or gender equality [101].

The health domain is also used by some studies, where it focuses on either the legalization of abortion (all studies that considered the SE16-T6 dataset), health insurance companies [133, 136], controversial health studies [67], or vaccination [96, 104]. Some other studies targeted scientific events, such as climate change (all studies that considered the SE16-T6 dataset) and natural disasters [118, 121, 124, 128].

4.3.3 Stance modeling

Generally, stance modeling can be performed at two levels: content and user levels. The content-level modeling includes textual and social media specific features (e.g., hashtags and mentions). The user-level modeling employs the user's network features, timeline, and profile information for stance detection. Figure 7 presents in detail the different forms of features at each level used for building stance detection models. Further, data on the features adopted in each of the 96 selected studies are listed in Tables 6, 7, and 8.

The majority of the studies in this SLR (87 out of 96) modeled the stance at the content-level by extracting one or more of the five feature levels: pragmatic, semantic, statistical, structural, and syntactic (see Fig. 7). Most studies extracted the semantic features of the text using static word embedding (e.g., Glove and word2vec) or contextual word embedding (e.g., Bidirectional Encoder Representations From Transformers (BERT)). Statistical features (e.g., N-gram) have also been widely employed to model the textual content, especially in earlier work (i.e., publications during 2015–2018). Pragmatic and syntactic features have been considered to model the textual content or enrich the textual content using external information, such as sentiment and emotion lexicons, target information, or syntactical dependency tree.

A graph-based approach was employed in four studies to perform a form of stance modeling at content-level [49, 63, 117, 133]. Wei et al. [117] proposed a modified graph convolutional network (GCN) to learn stance features by encoding conversation threads. Zhang et al. [49] used external emotion and semantic lexicons to build a semantic-emotion heterogeneous graph, which is then fed into a GCN to capture multi-hop semantic connections between emotion tags and words. Liang et al. [133] proposed an approach to capture the exact role of contextual words by investigating a novel technique of creating target-

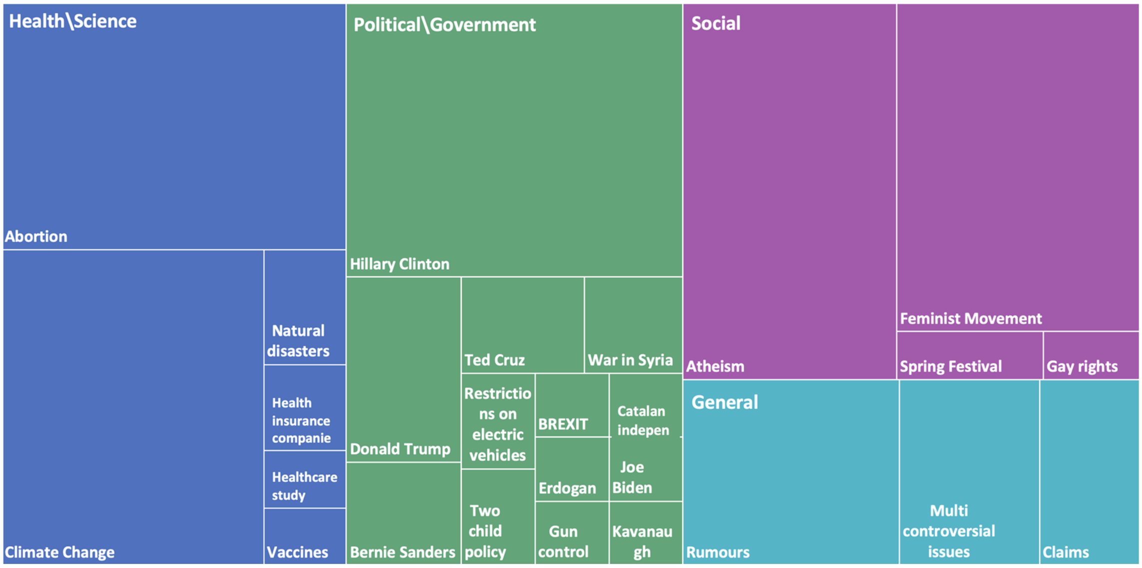

Fig. 6 Domain areas targeted by stance detection studies (sizes of rectangles represent the number of studies)

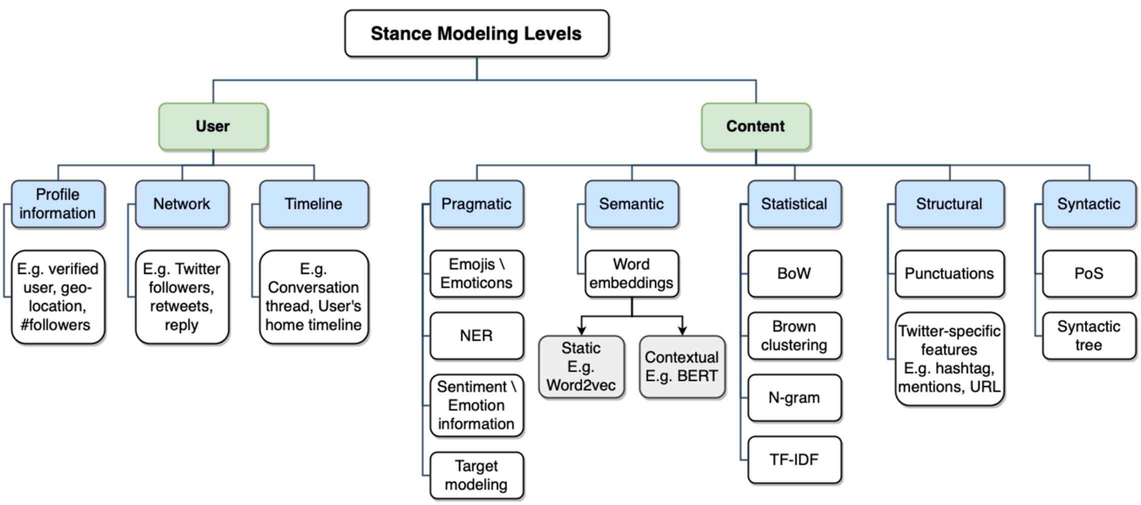

Fig. 7 Stance modeling

adaptive pragmatic dependency graphs with interactive GCN blocks for each tweet. Liu et al. [63] proposed a commonsense knowledge-enhanced model based on CompGCN [145]. The proposed model exploits both the semantic-level and structural-level information of the relation knowledge graph extracted from ConceptNet [65], allowing the model to improve its reasoning and generalization capabilities.

In addition, out of the 87 studies that modeled the stance at content-level, only one study [130] considered visual content with textual content. The authors of [130] proposed multimodal content as embedding vectors using BERT to obtain the embedding of the text content and used VGG19 to generate the visual embedding of an attached image.

User-level modality was used by a few stance detection studies (9 out of 96) compared with content-level modality. However, seven of the nine studies combined both users features and content features for stance detection [66, 68, 78, 90, 96, 119, 123], whereas two of them modeled the stance only at the user level [94, 95]. Darwish et al. [95] introduced a model for detecting the stance of prolific Twitter users using retweeted tweets, retweeting accounts, and hashtags as features for computing the similarities between users. Rashed et al. [94] used Google's convolutional neural network (CNN)-based multilingual universal sentence encoder to map the users into an n-dimensional embedding space. Furthermore, other studies [66, 68, 96] combined text embedding with user embedding generated from user network information comments such as likes,

Table 6 Supervised-based learning studies (ordered by: Type, Language, Dataset)

Type	Paper	Language	Features	ML models	Dataset	Best score (macro-F_1)
Target-specific	[48]	English	N-gram, SE, Sentiment lexicons	SVM	SE16-T6	59.21
	[70]	English	N-gram, Sentiment lexicons, Topic modeling	Maximum entropy	SE16-T6	61.04
	[71]	English	SE	CNN, voting scheme	SE16-T6	67.33
	[72]	English	N-gram, POS, Structural, Sentiment lexicons	LibSVM	SE16-T6	77.11
	[50]	English	SE, N-gram, POS, Sentiment lexicons, Target	SVM	SE16-T6	70.30
	[57]	English	Text, N-gram, Sentiment/subjectivity lexicons, Syntactic	SVM	SE16-T6	74.44
	[73]	English	SE, Target modeling	BiGRU, CNN	SE16-T6	67.40
	[76]	English	POS, Syntactic tree, Structural	SVM tree kernel, majority voting count	SE16-T6	70.03
	[77]	English	SE, Sentiment lexicon, Dependency parser, Argument information	LSTM+attention	SE16-T6	61.00
	[79]	English	SE, Target modeling	BiGRU+attention+memory network	SE16-T6	71.04
	[66]	English	Network, N-gram	SVM	SE16-T6	71.85
	[89]	English	SE	CNN+attention	SE16-T6	62.45
	[80]	English	SE, Target embedding	RNN-Capsule	SE16-T6	69.44
	[91]	English	TF-IDF, Sentiment lexicons	Weighted KNN	SE16-T6	76.45
	[92]	English	SE, POS, Structural, Statistical	Random forest, MLP, CNN, BiLSTM	SE16-T6	70.46
	[100]	English	CE, N-gram	Ensemble model (RoBERTa+BiLSTM+attention)	SE16-T6	73.77
	[106]	English	CE, Topic modeling, Sentiment, N-gram, TF-IDF	SVM, LR, Extremely Randomized Trees, AdaBoost	SE16-T6	74.63
	[51]	English	N-gram, Sentiment lexicons, Target modeling, Structural, SE	Ensemble classifier, DNRFAF, DSRFE, DECCV	SE16-T6, AM	SE16-T6: 71.24, AM: 57.61
	[35]	English	N-gram, BoW, Structural, Sentiment lexicons, Common-knowledge	SVM	TW-BREXIT	67.01
	[36]	English	SE, CE, Emotion lexicons	GRU, BERT	Procon20	76.90
	[90]	English	SE, User's timeline	LSTM+attention, GRU, Hierarchical LDA	Brexit, US Election-2016	Brexit:65, Election: 72
	[101]	English	SE, Temporal features	CNN	Temporally annotated	72.20
	[68]	English, Chinese	Network, SE, Topic modeling	CNN, LDA	CreateDebate, FBFans	75.50
	[75]	English, Chinese	SE, Target modeling	RNN, LSTM+attention	SE16-T6, NLPCC-2016	English: 68.79, Chinese: 72.88
	[93]	English, Chinese	SE	BiLSTM+attention	SE16-T6, NLPCC-2016	English: 69.21, Chinese: 74.14
	[55]	English, French, Italian, Spanish, Catalan	N-gram, BoW, Structural, Emotion lexicons, Domain knowledge	SVM, LR, CNN, LSTM, biLSTM	SE16-T6, IberEval 2017, Extended dataset for other languages	64.51
	[83]	Chinese	SE	CNN, GRU	NLPCC2016	62.20
	[82]	Italy	BoW, Structural	SVM	ConRef-STANCE-ita + user network	85.00
	[74]	Spanish, Catalan	BoW, POS, Structural	SVM	IberEval2017	Spanish: 48.88, Catalan: 49.01

Table 6 (continued)

Type	Paper	Language	Features	ML models	Dataset	Best score (macro-F_1)
Target-independent	[42]	English	SE, Structural, Text similarity	LSTM	RumourEval-17	43.40
	[108]	English	Structural, Sentiment score, POS, Text similarity	XGBoost	RumourEval-17	45.00
	[111]	English	CE, Conversation structure, Timestamp	CNN, BiGRU, MLP, attention	RumourEval-17	79.86
	[47]	English	Structural, Pragmatic, Conversation structure, Text similarity	SVM	RumourEval-17	47.00
	[119]	English	Structural, Similarity scores, Sentiment, User information	LR	RumourEval-17	57.40
	[112]	English	CE, Statistical, Structural, Sentiment lexicons	MLP, LSTM, GRU	FNC-1	83.08(Acc.)
	[109]	English	SE, Similarity score between claim and evidence	CNN, LSTM	FNC-1	56.88
	[114]	English	CE, SE, Structural, Statistical, Pragmatic, Text similarity, Sentiment, BLEU and ROUGE scores	BiLSTM+max-pooling+attention	FNC-1	82.23(Acc.)
	[129]	English	SE, Statistical, Sentiment, Text similarity, POS	Cascading classifiers, SVM, CNN	FNC-1	38.00
	[107]	English	BoW, Brown cluster, POS, Pragmatic, Structural, Confidence score, User profile information	Random forest	PHEME, RumourEval-17	PHEME: 77.42, RumourEval: 79.02(Acc.)
	[113]	English	SE, Structural, POS, BoW, Text similarity, Social network, Hawkes processes	LSTM-branch	PHEME	44.90
	[124]	English	CE, TF-IDF	RoBERTa, MLP	RumourEval-19	64.00
Cross-target	[49]	English	SE, Semantic/emotion lexicons, Knowledge graph	GCN, BiLSTM+knowledge-aware memory unit	SE16-T6	53.60
	[133]	English	CE, Syntactical dependency, Pragmatic dependency graph, Stance tokens	BiLSTM, GCN, attention	SE16-T6, WT-WT	SE16-T6: 59.5, WT-WT: 74.2
Multi-related targets	[142]	English	SE	Multi-kernel Convolution+Attentive LSTM	MultiTarget SD	58.72

retweets, mentions, and following accounts. Benton et al. [78] constructed user embeddings by combining textual embedding generated from Term Frequency-Inverse Document Frequency (TF-IDF), weighted bag of words (BoW), and social network embeddings. In addition, two studies utilized user profile information as a feature combined with textual features [119, 123]. Finally, the authors of [90] proposed to model the users' posts and the topical context of users' neighbors in social networks for user-level stance prediction.

The content-level-based approaches utilize the raw text for stance detection without the need for other information related to the writer. This feature makes these techniques applicable to all UGC platforms. However, relying solely on the text may not provide a complete understanding of the user's stance, especially when the user employs sarcasm to present his opinion on a specific topic. In contrast, user-level-based approaches employ the user's information to understand the user's stance. These approaches can be used only with UGC platforms that provide access to user information, such as social media.

To combine the features of content-level and user-level modelings, hybrid models have been proposed recently for stance detection. The attained results of these models outperformed the models that depend only on content-level. However, studies that depend on community features may compromise user privacy. This highlights the need for further research aimed at protecting social media users from unconsciously disclosing their views and beliefs. Therefore, due to privacy concerns, most social media platforms have recently begun restricting access to user information. This makes the application domain of user-level-based techniques limited and depends on the availability of the user's information.

4.4 Machine learning techniques (RQ4)

In this section, we consider RQ4. We analyze the ML approaches that contributed to the major developments in stance detection research. The ML techniques proposed for stance detection can be broadly classified into supervised-based, unsupervised-based, weakly supervised, and transfer

Table 7 Unsupervised-based and weakly supervised learning studies (ordered by: Type, Language, Dataset)

Type	Paper	Language	Features	ML models	Dataset	Best score (macro-F_1)
Target-specific	[84]	English	SE, Topic modeling, Noisy stance labeling	BiGRU, SRNet	SE16-T6	60.78
	[59]	English	N-gram, Emotion lexicons, Followers list	HL-MRFs, SVM	SE16-T6.B (unlabeled set)	57.52
	[78]	English	Network, BoW, TF-IDF	RNN+GRU	SE16-T6+ Tweets about gun control and gun rights	53.00
	[67]	English	TF-IDF, Predicted argument tags	SVM	1,063 comments about health study from news websites	77.00
	[96]	English	SE, CE, Network	Multilingual-BERT	Tweets on 8 polarizing US-centric topics	92.10
	[95]	English, Turkish	Network, Structural	UMAP, Mean shift, SVM	3 labeled sets (Kavanaugh, Trump, Erdogan), 1 unlabeled set of 6 topics in USA	90.40
	[94]	Turkish	CE, User's timeline	SVM, MUSE	108M Turkish election-related tweets+ Timeline tweets of 168k users	85.00
Target-independent	[53]	English	Text, Syntactical dependencies, Sentiment lexicons	Unsupervised approach	1,502 labeled arguments with consequences from Debatepedia	73.00
Cross- target	[131]	English	SE, CE, Target modeling	LSTM	SE16-T6	58.03

learning-based. The transfer learning models used for stance detection in turn can be subclassified into transductive, inductive, and low-shot. As highlighted in Fig. 4, 44 surveyed studies adopted supervised approaches for stance detection, five studies proposed weakly supervised models, four studies employed unsupervised models, and 43 studies applied transfer learning through unsupervised, supervised, or distantly supervised source tasks.

Figure 8 depicts a word cloud representing the frequency of ML techniques used in the selected studies. The significance of each technique is associated with its font size. It should be noted that the ML technique in this word cloud corresponds to the best reported technique (in terms of performance score) in each selected study. As shown in the figure, deep learning models that adopt the mechanism of self-attention (e.g., BERT) are used more frequently than the other approaches. The word cloud also shows that attention mechanism and recurrent neural network (RNN) models, such as long short-term memory (LSTM) and gated recurrent unit (GRU), are employed in a significant number of studies.

4.4.1 Supervised-based learning

The majority of the surveyed techniques (44 out of 96) applied supervised learning for stance detection. In supervised-based learning studies, the aim is to train a model on labeled data for a given target and domain, and expect it to perform well on new data of the same target and domain.

Earlier work in stance detection (between 2016 and 2018) employed traditional ML techniques to classify a stance toward a target. Most of them used a support vector machine (SVM) classifier [35, 47, 50, 66, 72, 74, 82]. Dey et al. [57] proposed a simple two-phase strategy with traditional SVM learning. A new syntactic feature was used in the first phase. This feature was learned from external subjectivity lexicons to differentiate between neutral and non-neutral tweets. In the second phase, non-neutral tweets were classified to favor or against by using a novel semantic feature extracted from external sentiment lexicons. Other stance detection studies employed other traditional ML techniques, such as random forest [107], gradient boosting [108], logistic regression (LR) [119], and k-nearest neighbors (KNN) [91]. A recent study [106], published in 2022, aimed to explain the stance detection model performance and provide a qualitative understanding of the classifier behavior. The authors exploited the Biterm Topic Model (BTM) to identify textual content that affected the stance. However, traditional ML techniques do not consider the contextual meaning of words. Given that having labeled data for every setting is infeasible, the performance score of such techniques is low compared with other approaches.

Several scholars have provided supervised models for stance detection using deep learning architectures. RNNs, a

Table 8 Transfer learning-based studies (ordered by: Type, Language, Dataset). **T:* Transductive, *S:* Sequential, *MT:* Multitask, and *LS:* Low-shot

Type	Paper	Language	Transfer learning type*				Features	ML models	Dataset	Best score (macro-F_1)
			T	S	MT	LS				
Target-specific	[69]	English		✓			SE, Hashtag prediction	LSTM	SE16-T6	67.80
	[85]	English			✓		SE, Topic modeling, Sentiment labeling	Attention	SE16-T6	68.54
	[54]	English			✓		SE, Sentiment and stance lexicons	Attention	SE16-T6	65.33
	[81]	English			✓		N-gram, BoW, Sentiment labeling	LSTM	SE16-T6	60.16
	[98]	English		✓			CE, Topic modeling	RoBERTa, Hierarchical capsule network	SE16-T6	78.43
	[99]	English		✓			CE, Target modeling	BERT+stance-wise convolution layer	SE16-T6	73.73
	[105]	English		✓			CE, Target modeling	BERTweet+AKD	SE16-T6, Multi-Target SD, AM, WT-WT, COVID-19, Election-2020	68.17
	[86]	English		✓			CE, Structural, Emotions	RoBERTa, LR	ACD, IAC 2.0	ACD: 77.13, IAC: 80.30
	[88]	English		✓			CE	LSTM, ULMFiT	ProCon	69.60
	[102]	English		✓	✓		CE	BERT+Adversarial attacks	10 datasets	66.95
	[103]	English		✓			CE, Stance relevant tokens	BERT	US election	77.27
	[104]	English					SE, N-gram, TF-IDF	BERT	7,530 labeled tweets about COVID-19 vaccination	78.94 (Acc.)
	[87]	English, Arabic	✓				CE	LSTM, CNN	FNC-1, Baly et al.	45.20
	[97]	Italy		✓			CE	UmBERTo	Sardistance	68.53
Target-independent	[38]	Arabic		✓			CE	Multilingual-BERT	Arabic News Stance	76.70
	[120]	English		✓			SE	BERT, Self-attention	RumourEval-17	47.50
	[117]	English			✓		CE, Stance information, Temporal modeling	BiGRU, Conversational-GCN, RNN	RumourEval-17	49.90
	[127]	English		✓	✓		CE, Conversation structure, Stance information	BERT, GCN	PHEME, RumourEval-17	PHEME: 42.70, RumourEval-17: 70.20
	[130]	English			✓		CE (textual: BistilBERT, visual: VGG-19)	Attention	PHEME, RumourEval-17	PHEME: 82.02, RumourEval-17: 80.41
	[122]	English		✓	✓		CE, Auxiliary data for paraphrase detection	BERT	FNC-1	74.40
	[110]	English			✓		SE, Stance information	GRU+enhanced shared-layer	FNC-1, PHEME	FNC-1: 32.80, PHEME: 43.00
	[123]	English			✓		CE, User profile information, Stance information	LSTM+task specific layers, VAE	PHEME	35.00
	[115]	English			✓		SE, Statistical(BoW, Brown clusters, Kullback-Leibler), Target modeling	Gaussian processes	PHEME, England Riots	PHEME: 59.80, England Riots: 70.80
	[121]	English		✓			CE, Structural, Cosine distance to source tweet	BERT, Ensemble methods	RumourEval-19	61.67
	[118]	English		✓			Structural, Sentiment information, User profile information	OpenAI GPT, input concatenation mechanism	RumourEval-19	61.87
	[128]	English		✓	✓		CE, SE, Structural, Pragmatic, Syntactic, Timeline	Longformer (trained from RoBERa), LSTM, Ensembling	RumourEval-19	67.20
	[125]	English		✓			CE	ALBERTv2	Args.me	73.70
	[116]	English		✓			CE	BERT+cosine embedding loss+joint loss	Perspectrum	79.95
	[126]	English		✓			CE, Confidence score, Negated perspective tokens	BERT	Perspectrum, IBM debater	Perspectrum: 81.35, IBM debater: 71.16

Table 8 (continued)

Type	Paper	Language	Transfer learning type* T	S	MT	LS	Features	ML models	Dataset	Best score (macro-F_1)
Cross-target	[132]	English	✓				SE, CE, Target modeling	CrossNet: Attention+MLP	SE16-T6	46.10
	[135]	English		✓		✓	CE, Topic modeling	Adversarial learning, 2-Layer feedforward network	SE16-T6	54.10
	[137]	English	✓				CE, Topic modeling, Sentiment	BERT, Adversarial attention network	SE16-T6, Perspectrum	68.47
	[138]	English		✓			CE, Entity recognition, Sentiment, Idelogy representation	RoBERTa, CNN	SE16-T6, VAST, Basil	67.66
	[139]	English			✓		CE, Syntactic, Sentiment, Opinion-toward	BERT, GAT, BiLSTM	SE16-T6, COVID-19, Election-2020	SE16-T6: 67.46, COVID-19: 82.6, Election: 79.37
	[33]	English		✓		✓	CE, Text similarity, Topic modeling	BERT, Ward hierarchical clustering	VAST	66.60
	[63]	English		✓		✓	CE, Knowledge graph	BERT, Concept-Net	VAST	70.20
	[64]	English				✓	CE, Knowledge graph, Sentiment, Commonsense representation	BERT, Graph autoencoder	VAST	72.6
	[134]	English	✓				CE, Target modeling	RoBERTa, Mixture-of-experts, Domain-adversarial learning	16 datasets	42.67
	[136]	English				✓	TF-IDF	MLP	WT-WT+ 134,922 synthetically annotated tweets	37.69
	[43]	French, German, Italian		✓		✓	CE	Multilingual-BERT	X-stance	76.60
Multi-related targets	[140]	English			✓		SE, Target modeling	BiLSTM, Dynamic memory	MultiTarget SD	56.73
	[141]	English			✓		SE	Seq2seq	MultiTarget SD	54.81
	[143]	English		✓	✓		CE, Target modeling	BERTweet, RoBERTa	MultiTarget SD, SE16-T6	60.56

robust class of artificial neural networks adapted to work for processing and identifying patterns in sequential data (e.g., natural language), were used in four studies [75, 80, 111, 114]. Sun et al. [80] were the first who introduced RNN-Capsule into the stance detection problem by extracting multiple vectors for stance features instead of one vector. In addition, they developed an attention mechanism to identify the dependency relationship between a text and a target. The results of experimenting with the proposed capsule network with attention mechanism on the SemEval-2016 dataset were encouraging with an average F1 score of 69.44.

LSTM, an RNN architecture specifically designed to handle long-term dependencies, is the most widely used deep learning architecture among the supervised learning studies used for stance detection (Fig. 8). LSTM algorithm was developed to deal with the vanishing gradient problem that can be encountered when training traditional RNN models. This feature makes the LSTM algorithm capable of learning and memorizing long-term dependencies. In total, 10 of 44 supervised learning studies made use of this algorithm [36, 42, 55, 75, 77, 90, 109, 112, 113, 142]. Five other studies [49, 92, 93, 100, 133] used BiLSTM, a variant of the standard LSTM that can improve the performance on sequence classification problems.

GRU is the newest entrant after LSTM and RNN; hence, it provides an improvement over them. Similar to LSTM, gates are used in GRU to control information flow. GRU is less complex than LSTM because it has a smaller number of gates. This advantage usually results in some performance improvements with GRU over LSTM. In total, 6 of 44 studies [36, 73, 79, 83, 90, 112] used GRU. Notably, Zhu et al. [90] introduced a novel neural dynamic model that jointly models topical contextual and user's sequential posting behavior. This model was simulated by a GRU to exploit the temporal contextual information for online learning. Their approach can dynamically identify topic-dependent stances, in contrast to static models that perform

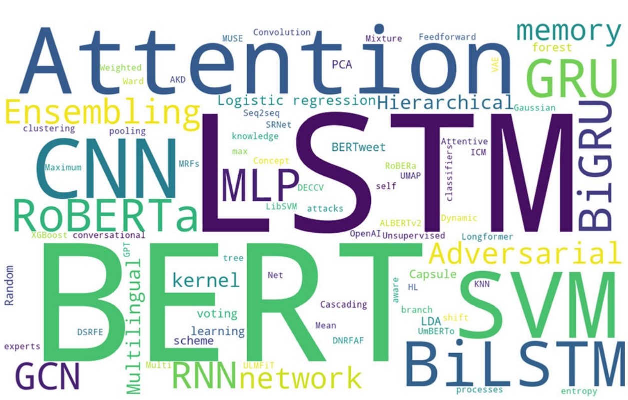

Fig. 8 A word cloud of the ML techniques used in the selected papers

one-time predictions. In addition, a two-channel CNN–GRU fusion network was proposed in [83] to overcome the problems of CNN, such as information loss when handling time-series data and not being able to extract features with varying lengths from text accurately.

Although LSTM and GRU are efficient for learning time-series data, these algorithms lack the capability of CNN for learning the spatial features of the input data [146]. To overcome this problem, several researchers employed CNN as a feature extractor and fed these features into a time-series learning technique, such as LSTM or GRU. CNN was used in 12 of the 44 studies [36, 55, 68, 71, 73, 83, 89, 92, 101, 109, 111, 129]. Notably, Mohtarami et al. [109] proposed a novel memory network model enhanced with LSTM and CNN networks based on a similarity-based matrix that has been used at inference time. Their results indicated that their model is capable of extracting significant snippets from an input text, which is useful not only for stance recognition but also for human experts deciding on the veracity of a claim.

Compared with single classifiers, ensemble classifiers have several advantages, such as decreasing the possibility of selecting an unstable subset of features, more precise prediction results, and avoiding the problem of local optimum [147]. Siddiqua et al. [76] proposed an ensemble learning approach in which different SVM tree kernel classifiers were consolidated to arrive at a final stance output using a majority voting scheme. In addition, the authors of [51] proposed an ensemble method using three algorithms—DNRFAF, DSRFE, and DECCV— developed in [148–150] for selecting the best set of features and classifiers. Chen et al. [100] proposed a novel ensemble model that combined a robustly optimized BERT approach (RoBERTa) with N-gram features and bidirectional LSTM (BiLSTM) with a target-specific attention mechanism. This fusion improved the results by 1.2% in micro-F1 score compared with state-of-the-art systems on the SemEval-2016 dataset. Similarly, Prakash et al. [124] used a multi-layer perceptron (MLP) with RoBERTa in an ensemble model with TF-IDF features as the input. This approach has been evaluated on the RumourEval-2019 dataset; the reported results indicated that the ensemble model outperformed the base RoBERTa model by 0.07% in the macro-F1 score, achieving state-of-the-art results [25].

For stance detection in a multilingual setting, Lai et al. [55] evaluated their model in four languages: English, Spanish, French and Italian. SemEval-2016 was used for English, IberEval-2017 for Spanish, and a new dataset was collected for French and Italian. Four types of features were used in this work: stylistic, structural, affective, and contextual. For feature learning and classification, several techniques were experimented with in this work, such as SVM, LR, CNN, LSTM, and BiLSTM. The reported results over all languages and domains of the classical ML models (SVM and LR) proved to be competitive compared with the considered deep learning models (CNN, LSTM, and BiLSTM). Three studies combined English and Chinese datasets to evaluate their models [68, 75, 93]. The authors of [68] proposed a CNN-based model by incorporating user information (from user comments and likes) and topic information obtained from topic modeling using linear discriminant analysis (LDA). Du et al. [75] proposed a neural attention model to extract target-related information for stance detection. Overfitting and gradient vanishing, as well as dealing with long-term dependencies during multilayer LSTM training, are all issues that were addressed in [93]. The authors presented a two-stage deep attention neural network that encodes tweet tokens with densely connected BiLSTM and target tokens with traditional BiLSTM.

Modeling of the interaction between stance and sentiment has been investigated by some researchers to boost the results of stance detection. Sobhani et al. [48] conducted several experiments to elucidate the interaction between stance and sentiment. They trained SVM using three features: N-gram, word embedding, and sentiment lexicon. They concluded that although sentiment features are useful, they alone are insufficient for stance detection. Ebrahim et al. [70] proposed maximum entropy (as discriminative) and Naive Bayes (as generative) to model the interactions between stance and sentiment by training the SemEval-2016 dataset. Hosseinia et al. [36] demonstrated that bidirectional transformers can achieve competitive performance, even without fine-tuning, by leveraging sentiment and emotion lexicons. Their findings suggested that employing sentiment information is more beneficial than emotion in detecting the stance.

The main advantages of supervised-based learning techniques are their reliable and accurate performance, given the appropriate representation of data and appropriate algorithms. However, the main drawback of these approaches is the need for a sufficient amount of annotated data for the desired task. Considering the plethora of human languages and the complexity of NLP problems in the real world, having labeled data for every setting is infeasible. Thus, supervised learning may fail given these real-world challenges.

Table 6 summarizes the supervised-based learning stance detection techniques used in the selected studies. In this table, we present the following comparison criteria:

- Type of the target dependency: target-specific, multi-related targets, cross-target, or target-independent.
- Target language of the study: single language or multilingual.
- Features used for model learning. The abbreviations in the feature column are SE: static embeddings, and CE: contextualized embeddings.
- ML models adopted by the study.
- Dataset name used for model training or the resource of collecting data if there is no defined dataset name.
- Best score. The literature on stance classification varies on the used performance measures; however, the macro-average F1 score is the most popular measure in the surveyed studies. Thus, we report the macro-average F1 score of the best ML model in each study. Note that few studies did not report their results in macro-average F1; thus, we report their results with the accuracy (Acc.) score.

As can be seen from Table 6, the majority of the supervised techniques have targeted a specific topic using the SemEval-2016 dataset. The English language is the main language considered by most supervised-based stance detection studies. In addition, deep learning models (e.g., LSTM and transformers) achieved higher performance scores compared to traditional ML models, such as SVM.

4.4.2 Unsupervised-based learning

Unsupervised learning has been used by a few studies in the field of stance detection (4 of 96; Table 7). Kobbe et al. [53] proposed an unsupervised approach for topic-independent stance classification. Their approach uses lexicons and grammatical dependencies to identify effective words in an argument and their impact. Rashed et al. [94] investigated the target-specific stance classification of Turkish political tweets. Their unsupervised approach involves mapping users into an n-dimensional embedding space using Google's CNN-based multilingual universal sentence encoder to represent the text of their tweets.

An unsupervised learning technique was proposed also in [96] to predict stance. The authors proposed an unsupervised clustering technique to predict a user's stance based on his/her timeline. This approach allows the model to automatically classify users with a few topical tweets with high accuracy (around 95%). Darwish et al. [95] proposed the use of unsupervised learning to tag numerous Twitter users with their stances on specific issues. They used different user features (e.g., retweeted users, vocabulary choices, and hashtags) as the basis for assessing user similarity. Then, the uniform manifold approximation and projection (UMAP) technique was employed for feature dimensionality reduction, followed by the mean-shift algorithm for user clustering. The hypothesis behind their approach is that users who share the same stance tend to communicate their opinions using the same vocabulary.

4.4.3 Weakly supervised learning

Several studies in this field attempted to employ weakly supervised learning for stance detection [59, 67, 78, 84, 131], which we present in Table 7. In this approach, the model learns from both labeled and unlabeled data. A simple weakly supervised learning approach is performed by setting a classifier from a small set of labeled data or heuristics and domain expertise and using the classifier to estimate labels for the unlabeled data. The unlabeled data predictions can be used as "pseudo-labels" by adding them to the training set. Weakly supervised learning is a powerful approach for solving problems that require a large amount of data with little supervision. Although this method addresses the issue of a lack of labeled data, it is not as accurate as supervised learning. Moreover, the long training time and poor generalization are the main limitations of this learning approach.

Some studies automatically annotated data by employing a rule-based classifier (e.g., using regular expressions) [78, 131]. Ebrahimi et al. [59] were the first to employ statistical relational learning for stance detection. They used hinge-loss Markov random fields (HL-MRFs) to constrain pairs of similar tweets and pairs of neighboring users to have similar labels. Sobhani et al. [67] are pioneers in stance classification from the NLP perspective; they proposed a framework for stance classification at the document level based on topic modeling (nonnegative matrix factorization). The main advantage of the proposed framework is that it is minimally supervised, as it does not require labeled data. They collected 781 comments from news websites and annotated them with a predefined list of arguments resulting from topic modeling. The linear SVM was used to classify the stance based on the TF-IDF features and the predicted argument tags.

Although distant-supervision approaches have been developed to alleviate the difficulty of manually annotating huge amounts of training data, they are hindered by the problem of noisy labeling. Thus, Wei et al. [84] proposed a reinforced technique comprising two models. The first model is a topic-aware detection network for topic learning, and the second is a stance revision policy network that learns to eliminate noisy labeling based on off-policy reinforcement learning.

4.4.4 Transfer learning based

In the field of ML, transfer learning is a well-known method to leverage unlabeled data in the source domain or in the target domain to the most effect [151]. Thus, transfer learning is essentially a semi-supervised technique with domain adaptation. Broadly, transfer learning can be defined as the process of training a model on a large-scale dataset and then using that pre-trained model to learn for a target task [152]. Recently, pre-trained language models, such as OpenAI GPT [153], Google AI's BERT [154], ELMo [155], and ELECTRA [156], have revolutionized the field of transfer learning in NLP. Many scholars have adapted transfer learning for stance detection, mainly from the NLP perspective. Thus, in the following, we introduce definitions and taxonomies of transfer learning that are most commonly encountered in stance detection studies.

Transfer learning is defined as the technique used to transfer knowledge from related tasks, domains, and languages for various scenarios [157]. The different scenarios of NLP problems lead to the definition of a taxonomy for transfer learning, specifically for NLP. A taxonomy provided by Ruder [151] divides the scenarios based on the source task and the target task. The situation when we have a source task different from the target task is defined as *inductive transfer learning*, whereas the *transductive transfer learning* is used when the source and target tasks are the same. In addition, a recent study area of transfer learning is *low-shot learning*, which is introduced to train a model for a task with a small amount of labeled data. We follow this taxonomy in categorizing the transfer learning techniques used in the selected studies.

Table 8 presents the stance detection studies that proposed transfer learning based models. In the following, we introduce the reader to the techniques adopted in the surveyed literature, following the three classes of transfer learning: (A) transductive transfer learning, (B) inductive transfer learning, and (C) low-shot learning.

4.4.4.1 Transductive transfer learning Transductive transfer learning is generally considered when there is sufficient labeled data in the source domain only and when the aim is to learn representations that are useful for a specific target domain rather than being beneficial in general. This type of transfer learning is useful for real-world problems where the distribution of the test data differs from the training data.

Transductive transfer learning is employed in the reviewed studies for domain adaptation [132, 134, 137] and for cross-lingual learning [87]. In cross-lingual learning, the documents in the source and target domains are written in two different languages; hence, the feature spaces differ. By contrast, the documents in domain adaptation are written in the same language but from different domains or about different targets (e.g., source documents about political tweets and target documents of tweets about social issues). However, the main problem in domain adaptation is negative transfer [151]. This problem usually results from dissimilarity between domains. Therefore, most of the approaches targeting domain adaptation have attempted to mitigate this problem.

Transductive transfer learning was proposed in [134] to learn out-of-domain prediction of unseen targets. An end-to-end system was proposed for learning from heterogeneous labels based on label embeddings and unsupervised domain adaption as well as an unsupervised method for predicting an unseen set of user-defined targets based on label name similarity.

Xu et al. [132] studied the potential for generalizing classifiers between different domain-related targets. A novel self-attention neural model was proposed to extract target-independent information. The proposed model can transfer knowledge from a source target to a destination target and outperformed several baselines in some domains, according to experimental results. Similarly, Sun et al. [137] investigated the possibility of bridging the gap between different target data by proposing an adversarial attention network. The proposed model learn the correlation of the posts from different targets by determining and connecting the sentiment and the topic information of each post.

Mohtarami et al. [87] were the first to introduce a model for cross-lingual stance detection. They developed an end-to-end feature-light memory network based on contrastive stance alignment. This network aligns the source and target languages' class labels for an effective language adaptation. They conducted the experiments on English (as the source language) using the Fake News Challenge dataset (FNC-1) [30] and Arabic (as the target language) using the Arabic benchmark dataset [37]. Their proposed method can address the challenge of limited labeled data in the target language. However, there is a large room for improvement since their model achieved an F1 score of 45.2.

4.4.4.2 Inductive transfer learning Inductive transfer learning improves the performance of the target task using the knowledge learned from the source task. This type of learning is distinguished from transductive learning by the fact that it can be applied between different tasks [151]. In inductive transfer learning, there is a distinction between multitask learning (MTL) and sequential transfer learning (STL). *MTL* is a learning paradigm that aims to leverage useful information contained in related tasks simultaneously to enable a model to generalize better on the target task. Whereas in *STL*, models learn tasks sequentially rather than simultaneously. In other words, in STL, models learn each task separately and not jointly optimized as in MTL.

In the following, we illustrate the taxonomy of inductive transfer learning more using one example scenario of stance detection:

1. Consider having two source tasks: "language modeling" and "sentiment classification", and a target task: "stance detection". Language modeling is a task based on unlabeled data, whereas sentiment classification and stance detection are tasks based on labeled data. Since we have different tasks, we will follow the inductive transfer learning approach.
2. If we are using language modeling and the labeled data to learn the two other tasks (sentiment classification and stance detection) simultaneously, we are following *MTL*.
3. If we are using language modeling and labeled data to learn sentiment classification first and later use this knowledge to learn stance detection, we are following *STL*.

Among the selected 96 studies, 33 proposed models using inductive transfer learning techniques; particularly, 17 followed STL, 11 applied MTL, and 5 employed both approaches. We summarize those that employed the STL and MTL techniques in the following sections.

Sequential transfer learning

In NLP, STL is arguably the most commonly used type of transfer learning [151]. From the definition of STL presented above, the goal is to transfer knowledge from the source task model to improve the target model's performance. Although STL is a time-consuming technique during source model training, it quickly adapts to the target task. The reviewed studies present different scenarios of STL using unsupervised source tasks [38, 88, 98, 99, 102–105, 116, 118, 121, 125–128, 138, 143], supervised source tasks [86, 120, 122], and distantly supervised source tasks [69, 97].

Unsupervised STL (also called unsupervised pretraining) is the most common scenario used in the reviewed studies. It allows a model to capture more general characteristics of language structure and meaning, making it more transportable. Most unsupervised pretraining techniques focus on learning contextual representations of words from large unlabeled data, which is done by having an entire network that is pre-trained in an unsupervised approach with a language modeling objective, and then the model is fine-tuned on the classification task. Most reviewed studies used language models, such as BERT [154] used in [38, 99, 102–104, 116, 121, 126, 127, 143], OpenAI GPT [153] used in [118], or RoBERTa [158] used in [98, 125, 128, 138].

Notably, Zhao and Yang [98] proposed a novel approach by applying a pre-trained RoBERTa model [158] with a hierarchical capsule network. They combined the relevant topic information with each tweet and used a related textual entailment task for fine-tuning. The evaluation results on the SemEval-2016 dataset indicated that the proposed model significantly improved the performance by 6.32% in average F1 score compared with the first-place state-of-the-art model. In addition, their findings suggested that using a pre-trained language model directly with only a fully connected layer (without the hierarchical capsule network) would lose meaningful information in texts. For the political domain, Liu et al. [138] provided a new large language model (called POLITICS) that is generated by continuing training RoBERTa on a large-scale dataset comprising political news articles. Using ideology-driven pretraining objectives in the training process, POLITICS provides a general-purpose method of analyzing ideological content.

Hosseinia et al. [88] established a dataset from *ProCon.org*, comprising a collection of controversial issues. They proposed a model inspired by ULMFiT [159], which is a framework for pretraining and adapting learned representations. The proposed model comprises three units. The first unit is a parallel language model unit for learning the argument and context of the target. The other units are a fusion unit to summarize all data elements, and a classification unit to classify the stance. In their analysis, they showed that the dataset is challenging, but fine-tuning the pre-trained language model on context information yields a competitive performance.

The study by Khouja [38] is the only study that investigated stance detection for the Arabic language, specifically, the target-independent stance detection for claim verification. Khouja [38] established an Arabic corpus comprising news headlines, which was modified into a new claim by annotators; thus, the dataset comprised pairs (claim, evidence). LSTM and multilingual BERT were explored and developed to build a baseline for claim-based stance detection for Arabic. The proposed baseline model achieved an F1 score of 76.7.

Instead of using transformers (e.g., BERT) to encode the contextual representation of texts in unsupervised settings, Bugueno et al. [120] used the output of a set of supervised baseline techniques for a transformer. The outputs of the baselines were combined with the texts to generate an encoding of the baselines' outcomes. Then, the transformer-proportioned attention matrices were used to determine relevant baselines for the model. Another supervised pretraining approach, in [86], extracts the context of a debate by looking for feasible combinations of pairs of posts specific to each topic. The authors followed the flow of the dialogue and learned the language inference between phrases to establish the stance class while respecting the timestamps of each sentence. They generated features with RoBERTa (for the sentence-pair classification) and then trained a secondary classifier to map each sentence onto the set {Agreement, Disagreement, Neutral}.

For the distantly supervised pretraining settings, Zarrella et al. [69] used an RNN initialized with features learned from two large unlabeled datasets via distant supervision. Using the features, they exploited a hashtag prediction auxiliary task to learn post representations, which were fine-tuned on several hundred labeled instances for stance detection. Their model achieved the best performing system for SE16-T6-A [21]. In a more recent study [97], the authors examined the potential contribution of three auxiliary tasks: sentiment, irony, and hate speech detections. They fine-tuned Italian BERT language modeling [154] and augmented each input in the training data with labels of the three auxiliary tasks. Their system achieved the best performing system in the Sardistance competition [26].

Multitask learning

MTL contributed to machine learning success in various applications [160]. MTL (also called joint learning) aims to improve the generalization of a model on a target task by deriving knowledge from the training signals of auxiliary tasks. The related tasks in MTL create an "inductive bias", causing the model to favor hypotheses that can explain more than one task. As presented in Figure 4, 11 studies (out of 96) applied MTL [54, 81, 85, 110, 115, 117, 123, 130, 139–141], and five studies implemented both MTL and STL [102, 122, 127, 128, 143].

In cases where we want to get predictions for multiple tasks at once, MTL is a natural fit. Fang et al. [122] were the first to apply MTL to the problem of stance detection using multiple NLP-related tasks (i.e., sentiment analysis, paraphrase detection, question answering, and textual entailment). The resulting model of both unsupervised and supervised pretraining on these tasks was fine-tuned on the target stance detection task. Their proposed MTL model outperformed state-of-the-art systems by 14.4% in macro-F1 score on FNC-1 [30].

Four studies integrated stance and sentiment detection jointly via MTL [54, 81, 85, 139]. Sun et al. [81] argued that using a feature-based discrete model cannot efficiently handle the interaction between stance and sentiment. Thus, they proposed a joint neural model based on LSTM to integrate the sentiment features. Similarly, the authors of [54] proposed a joint neural model to integrate both sentiment attention and target attention. The loss function of the proposed model used existing sentiment and stance lexicons to guide the attention mechanism. The proposed model significantly improved the performance on the SemEval-2016 dataset. Chauhan et al. [85] leveraged the interdependence of stance and sentiment via a multitask deep neural model and developed an effective attention-based technique that integrated contributing features by setting more attention to the relevant words in a post. Lastly, a recent study by Fu et al. [139] argues that relying on sentiment information alone for stance detection is not sufficient, since authors' opinions may be toward a target or toward other aspects. Thus, they developed an MTL model using a label relation matrix that considers *opinion-*

toward classification and *sentiment classification* as auxiliary tasks for the main task (i.e., stance detection).

Seven studies proposed an MTL framework to tackle stance detection and rumor veracity prediction jointly [110, 115, 117, 123, 127, 128, 130]. Notably, Zhang et al. [130] proposed an MTL model that shared higher meta-network layers to capture the meta-knowledge of textual and visual contents. Each task (i.e., stance detection or rumor veracity) benefited from the shared meta-knowledge by dynamically producing the parameters of task-specific models. This method is opposed to generic MTL approaches that share lower network layers to extract common features.

Three studies [102, 140, 141] proposed deep learning models trained on multiple targets in a multitask fashion, such that detecting stances toward N targets was regarded as a set of N tasks. However, unlike the previously presented studies that showed that applying MTL techniques improves the model performance, the reported results in Sobhani et al. [141] indicated that their proposed single-task attention-based model is more effective than the multitask LSTM model on the Multi-Target SD dataset [28].

4.4.4.3 Low-shot learning ML techniques have proven to be quite effective in NLP tasks and data-intensive applications in general; however, they struggle when the training dataset is limited. Recently, low-shot learning has been proposed as a solution to this problem. The goal of this learning paradigm is to generalize to new tasks that have limited training data (zero or few labeled examples) using prior knowledge [161]. In particular, training ML models when just a few examples with supervised information are provided is called *few-shot learning*, whereas *zero-shot learning* attempts to predict the correct class without being exposed to any examples with supervised information for that class. The lack of labeled samples makes the estimation of the loss value during model training more challenging, which is the key issue of few-shot learning.

Six studies (of the 96) proposed low-shot classifiers [33, 43, 63, 64, 135, 136]. All of the six low-shot models were proposed for cross-target stance detection (Table 8). This indicates the need for low-shot techniques to improve the generalization across topics [162].

Allaway et al. [33] were the first who introduced low-shot learning for stance detection. In particular, they developed a new dataset, called VAST, comprising thousands of topics covering broad themes. VAST was proposed to fill the gap of existing datasets that contain a limited number of topics (e.g., five topics) and to evaluate generalization when we have only a few examples per topic. Using this training dataset, they proposed a new stance detection approach that uses generalized topic representations to implicitly capture links between topics.

VAST was also adopted in [63], where the authors presented a commonsense knowledge-enhanced module to exploit both the semantic-level and structural-level information, allowing the model to improve its reasoning and generalization capabilities. Nevertheless, in their model, knowledge is restricted to knowledge relationships between documents and topics. To boost the transferability of knowledge, Luo et al. [64] proposed a model that includes, besides the commonsense knowledge-enhanced module, a graph autoencoder module to obtain other types of commonsense information. Their model achieves state-of-the-art performance on the VAST dataset.

The authors of [135] introduced a zero-shot model that uses adversarial learning, following the success of the domain-transfer architecture by [163], to produce topic-invariant representations allowing the model to generalize to unseen topics. Conforti et al. [136] proposed the use of synthetically annotated data and a weakly supervised framework to improve cross-target generalization.

Unlike the previously presented studies that proposed low-shot models for generalization across topics (i.e., cross-target), Vamvas et al. [43] proposed a zero-shot model for generalizing across languages (i.e., cross-lingual), aside from the generalization across topics. They fine-tuned multilingual BERT on a new dataset comprising French, German, and Italian comments on political issues, allowing for a cross-lingual and cross-target evaluation of stance detection.

4.5 Research gaps (RQ5)

In this section, we aim to answer RQ5 by presenting the research gaps and promising future trends in the stance detection field. We analyzed the surveyed studies and found that there are still many limitations in the previous research work that could provide a pathway to future research. The identified gaps are as follows:

- *Complexity of the model:* We found that stance detection studies lean in either one of the following approaches: (1) a complex representation model that uses numerous manually crafted features to improve the learning process using human judgment and (2) an excessively simple feature model that is built only on raw term frequencies and fed to a complex classifier. However, both of these approaches have limitations. Complex feature models are highly domain-specific and may be impractical. Furthermore, studies that depend on community features may compromise user privacy. Meanwhile, models that depend on raw term frequencies fed into complex classifiers are turned into black boxes in which it is impossible to explain their performance.

The models' explainability does not seem not to be a research priority in the surveyed literature. However, the lack of explainability is a necessary concern in the practical implementation of stance detection models in areas such as polling predictions for referendums and elections, online public health surveillance, and trend and market analysis due to the possibility of inaccurate predictions. Further, when stance detection is employed as a module for detecting fake news, the model's explainability may be critical.

- *Language:* Despite the growing interest in studying stance classification, only 13% of the selected studies analyzed contents not written in English. These studies include [68, 75, 83, 93] in Chinese, [94, 95] in Turkish, [82, 97] in Italian, [38] in Arabic, and [43, 55, 74] considering multiple languages (i.e., Spanish, Catalan, French, and German). Some languages pose many challenges, e.g., the semantic analysis of Arabic text is particularly difficult due to its rich and complex morphology, orthographic ambiguity, orthographic noise, and dialectal variations [95, 164]. Furthermore, current research has a language orientation; a model that is independent of language could be revolutionary in this research area.
- *Resources:* As observed from Table 3, most available datasets are in the English language. We noticed that many studies (including the recent publications in 2022) are still on the old public dataset, SemEval-2016, for stance detection or related applications. We believe that this field requires more benchmarked datasets to be published under a common open license for public use. Any of the following criteria should be targeted: non-English, multilingual data, and annotations of different opinion dimensions (e.g., emotion, sarcasm, and irony).

  Furthermore, the manual annotation of data by crowdsourcing services is a typical approach currently used; nevertheless, this strategy can introduce annotators' bias into the data. Thus, non-intrusive data collection strategies need to be investigated by researchers.

  In addition, most pre-processing tools and resources (presented in Sect. 4.1.3) only support English. For example, many scholars have created stance detection models using sentiment and emotion lexicons (e.g., VADER and NRC). However, these resources are limited to the English language. Future studies can develop such tools for non-English languages to support stance detection and sentiment analysis models.
- *Reproducibility:* In this survey, we found only one study that performed a systematic comparison of stance classification methods, which was conducted by Ghosh et al. [165]. They investigated seven target-specific stance detection models through experiments on two datasets: SE16-T6, and a formal text dataset of health-related articles. This study highlighted the challenges in the reproducibility of the experimented stance detection models. The evidence from this study indicates that there is no single model that can provide a satisfying metric value for all datasets.
- *Sentiment and sarcasm features:* Some studies found a great interaction between stance and sentiment [36, 54, 70, 81], whereas others demonstrated that it is inefficient to use a sentiment as a feature for stance detection models [48, 50, 166]. Thus, hypotheses regarding the interaction between sentiment and stance appear to be ill-defined and debatable. Regarding the sarcasm feature, as observed from [165], the errors were mostly in texts that contained sarcastic comments. Thus, analysis of the interaction between sarcasm and stance could benefit these methods. We did not find any study that considered sarcasm features for stance detection. In addition, no study takes into account all of the numerous social dimensions in their research, such as emotions, sentiment, and sarcasm.

  Furthermore, MTL can bring improvement in the performance of many machine learning techniques [160, 163, 167], which is observed also in our literature review presented in Sect. 4.4.4. Thus, studying and evaluating a joint neural architecture based on the MTL paradigm that jointly models related social dimensions should be investigated further.
- *Diachronic evolution:* Given that people's opinions might change over time [17], recording and evaluating temporal data is essential in studying stance evolution. The goal of studying diachronic evolution is to understand the temporal variations in the real world and their impact on public opinion. Diachronic evolution analysis also allows the identification of factors influencing people's stances. The evolution of the users' stance can be better analyzed with a model that incorporates context from multiple time periods. Despite this, state-of-the-art studies do not look in this research direction. From the analysis of this SLR, only three studies [35, 82, 101] considered a diachronic aspect to elucidate users' stance dynamics. However, the research on the diachronic evolution of stance is still in its early stage with several aspects that have yet to be investigated, also attributable to a scarcity of large datasets collected over long periods of time.
- *Modality:* Another sub-domain that demands further research is *multimodal* stance detection. The current research focuses solely on the text modality; however, there are opportunities in combining textual modality with other modalities, including visual (e.g., videos and

images) and audio, to analyze how these data perform together. In this SLR, there is only one study [130] that proposed multimodal learning considering visual content aside from the textual content, which achieved state-of-the-art performance on two Twitter benchmark datasets.

- *General stance classifier:* Although many of the proposed models have achieved excellent performance in stance detection, they present crucial flaws. First, the proposed models cannot effectively identify the relationship between the target and text, which plays a key role in stance detection. Modeling the dependency relationship between the target and text could improve the performance of stance detection. Second, most current techniques for stance detection use topic-based learning (i.e., the target is defined and annotated for the model). Adopting nontopical aspects to the current techniques has not been sufficiently explored in the literature for stance detection. The current models need further enhancement to adapt to the targets of interest without the need for annotated data for each target. This might lead to a general stance classifier that has comparable performance with supervised target-specific stance detection.

5 Conclusion

This SLR was conducted on stance detection research, which totals 96 published studies, selected by a filtering process of 1216 studies from six databases, and spans a period of seven years between 2015 and 2022. We performed a full reading of these publications to address five research questions. Through this SLR, we provided in-depth analysis and insights into the types of ML techniques, comparison between the proposed models in terms of performance score, datasets and resources used, domains and application areas, and other aspects derived. We proposed a taxonomy that allows studies to be grouped into different dimensions so that similarities and differences between approaches may be observed. A mapping of experiment settings was also a part of this SLR, which we hoped would aid in the design of new studies.

Our final discussion on the SLR listed the gaps to be explored for future research toward more robust approaches for stance detection. Potential future directions in this area include developing a more realistic and holistic framework for explaining how stance detection models work. Regarding language orientation, future stance detection studies need to consider cross-lingual and multilingual approaches. In addition, language-independent models could be revolutionary in this field. In terms of resources, there is a need for establishing new datasets that consider any of the following criteria: non-English, multilingual, and annotations of different opinion dimensions. Furthermore, current methods need to pay more attention to integrating external knowledge of different opinion dimensions, and incorporating non-textual modalities (e.g., videos and images). In addition, incorporating temporal data to study the diachronic evolution of stance is still in its early stage and needs to be further examined. Lastly, current models need to be enhanced to fit the targets of interest without requiring annotated data for each target, which could result in a general stance classifier that is comparable to supervised target-specific classifiers.

Although we believe that this SLR has useful information regarding stance detection research, there are still some limitations that may affect the scope. The procedure of finding all relevant studies and selecting digital search libraries is a common threat to SLR [168]. To address this threat, six well-known digital databases were selected and thoroughly examined: ACM, Scopus, Springer, Web of Science, IEEE-Xplore, and Google Scholar. In addition, we manually defined the search string based on related review studies to reduce bias. Another limitation is that the more recent studies are not included in this SLR due to the time involved in analyzing the review corpus to obtain credible results; therefore, forward snowballing may provide improvements, as we only performed backward snowballing.

Acknowledgments The authors would like to acknowledge the support received from the Saudi Data and AI Authority (SDAIA) and King Fahd University of Petroleum and Minerals (KFUPM) under the SDAIA-KFUPM Joint Research Center for Artificial Intelligence Grant JRC-AI-RFP-05.

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Nguyen D, Doğruöz AS, Rosé CP, de Jong F (2016) Computational sociolinguistics: a survey. Assoc Comput Linguist 42:537–593. https://doi.org/10.1162/COLI
2. Küçük D, Fazli CAN (2020) Stance detection: a survey. ACM Comput Surv. https://doi.org/10.1145/3369026
3. Bois JWD (2007) The stance triangle. Stancetaking Discourse: Subj Eval Interact 164:139–182
4. Kockelman P (2004) Stance and subjectivity. J Linguist Anthropol 14:127–150
5. Jaffe A et al (2009) Stance: Sociolinguistic Perspectives. Oxford University Press, US

6. Grimminger L, Klinger R (2021) Hate towards the political opponent: a twitter corpus study of the 2020 us elections on the basis of offensive speech and stance detection. arXiv
7. AlDayel A, Magdy W (2021) Stance detection on social media: state of the art and trends. Inform Process Manag. https://doi.org/10.1016/j.ipm.2021.102597
8. Kratzwald B, Ilić S, Kraus M, Feuerriegel S, Prendinger H (2018) Deep learning for affective computing: text-based emotion recognition in decision support. Decis Support Syst 115:24–35. https://doi.org/10.1016/j.dss.2018.09.002
9. Kumar A, Narapareddy VT, Srikanth VA, Malapati A, Neti LBM (2020) Sarcasm detection using multi-head attention based bidirectional lstm. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2963630
10. Lin J, Mao W, Zeng D (2016) Competitive perspective identification via topic based refinement for online documents. In: IEEE international conference on intelligence and security informatics: cybersecurity and big data, ISI 2016. https://doi.org/10.1109/ISI.2016.7745474
11. Stab C, Miller T, Schiller B, Rai P, Gurevych I (2018) Cross-topic argument mining from heterogeneous sources using attention-based neural networks. In: Proceedings of the 2018 conference on empirical methods in natural language processing, EMNLP 2018. https://doi.org/10.18653/v1/d18-1402
12. Coletto M, Garimella K, Gionis A, Lucchese C (2017) Automatic controversy detection in social media: a content-independent motif-based approach. Online Soc Netw Media. https://doi.org/10.1016/j.osnem.2017.10.001
13. Hube C, Fetahu B (2019) Neural based statement classification for biased language. In: WSDM 2019 - Proceedings of the 12th ACM international conference on web search and data mining. https://doi.org/10.1145/3289600.3291018
14. Cortis K, Davis B (2021) Over a decade of social opinion mining: a systematic review. Artif Intell Rev 54(7):4873–4965
15. Jesson J, Matheson L, Lacey FM (2011) Doing your systematic review - traditional and systematic techniques vol 3,
16. Hardalov M, Arora A, Nakov P, Augenstein I (2021) A survey on stance detection for mis- and disinformation identification. arXiv preprint, 1–9
17. Alkhalifa R, Zubiaga A (2021) Capturing stance dynamics in social media: open challenges and research directions. arXiv
18. Wang R, Zhou D, Jiang M, Si J, Yang Y (2019) A survey on opinion mining: from stance to product aspect. IEEE Access 7:41101–41124. https://doi.org/10.1109/ACCESS.2019.2906754
19. Kitchenham B (2004) Procedures for performing systematic reviews. Keele University, UK and National ICT Australia vol 33, pp 1–26. https://doi.org/10.1.1.122.3308
20. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Group, T.P (2009) Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. Ann Internal Med 151:264–269. https://doi.org/10.1371/journal.pmed.1000097
21. Mohammad SM, Kiritchenko S, Sobhani P, Zhu X, Cherry C (2016) Semeval-2016 task 6: detecting stance in tweets. 10th International Workshop on Semantic Evaluation (SemEval-2016), pp 31–41. https://doi.org/10.18653/v1/s16-1003
22. Xu R, Zhou Y, Wu D, Gui L, Du J, Xue Y (2016) Overview of nlpcc shared task 4: stance detection in chinese microblogs. Natural language understanding and intelligent applications, pp 907–916. https://doi.org/10.1007/978-3-319-50496-4_85
23. Taulé M, Martín MA, Rangel F, Rosso P, Bosco C, Patti V (2017) Overview of the task on stance and gender detection in tweets on catalan independence at ibereval 2017. In: 2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017, vol 1881, pp 157–177
24. Derczynski L, Bontcheva K, Liakata M, Procter R, Hoi GWS, Zubiaga A (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), pp 69–76
25. Gorrell G, Kochkina E, Liakata M, Aker A, Zubiaga A, Bontcheva K, Derczynski L (2019) Rumoureval 2019: determining rumour veracity and support for rumours. In: Proceedings of the 13th international workshop on semantic evaluation (SemEval-2019), pp 845–854
26. Cignarella AT, Lai M, Bosco C, Patti V, Rosso P (2020) Sardistance @ evalita2020: overview of the task on stance detection in italian tweets. EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, vol 2765, pp 1–10. https://doi.org/10.4000/books.aaccademia.7084
27. Ferreira W, Vlachos A (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. ACL, pp 1163–1168
28. Sobhani P, Inkpen D, Zhu X (2017) A dataset for multi-target stance detection. In: 15th conference of the European chapter of the association for computational linguistics, EACL 2017, vol 2, pp 551–557. https://doi.org/10.18653/v1/e17-2088
29. Bar-Haim R, Bhattacharya I, Dinuzzo F, Saha A, Slonim N (2017) Stance classification of context-dependent claims. In: Proceedings of the 15th Conference of the European chapter of the association for computational linguistics, Volume 1, Long Papers, vol 1, pp 251–261 (2017)
30. Hanselowski A, Schiller B, Caspelherr F, Chaudhuri D, Meyer CM, Gurevych I (2018) A retrospective analysis of the fake news challenge stance-detection task. In: Proceedings of the 27th international conference on computational linguistics (COLING 2018)
31. Chen S, Khashabi D, Yin W, Callison-Burch C, Roth D (2019) Seeing things from a different angle: discovering diverse perspectives about claims. In: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 1. https://doi.org/10.18653/v1/n19-1053
32. Ajjour Y, Wachsmuth H, Kiesel J, Potthast M, Hagen M, Stein B (2019) Data acquisition for argument search: The args.me corpus. Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz), vol 11793 LNAI, pp 48–59. https://doi.org/10.1007/978-3-030-30179-8_4
33. Allaway E, Mckeown K (2020) Zero-shot stance detection: a dataset and model using generalized topic representations. In: Proceedings of the 2020 conference on empirical methods in natural language processing, pp 8913–8931
34. Conforti C, Berndt J, Pilehvar MT, Giannitsarou C, Toxvaerd F, Collier N (2020) Will-they-won't-they: a very large dataset for stance detection on twitter. arXiv. https://doi.org/10.18653/v1/2020.acl-main.157
35. Lai M, Patti V, Ruffo G, Rosso P (2020) Brexit: Leave or remain? The role of user's community and diachronic evolution on stance detection. J Intell Fuzzy Syst 39:2341–2352. https://doi.org/10.3233/JIFS-179895
36. Hosseinia M, Dragut E, Mukherjee A (2020) Stance prediction for contemporary issues: Data and experiments. In: Proceedings of the eighth international workshop on natural language processing for social media, pp 32–40. Association for Computational Linguistics (ACL), Online. https://doi.org/10.18653/v1/2020.socialnlp-1.5
37. Baly R, Mohtarami M, Glass J, Moschitti A, Nakov P (2018) Integrating stance detection and fact checking in a unified corpus. arXiv

38. Khouja J (2020) Stance prediction and claim verification: an Arabic perspective. Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER), pp 8–17
39. Lai M, Patti V, Ruffo G, Rosso P (2018) Stance evolution and twitter interactions in an Italian political debate. In: International conference on applications of natural language to information systems, vol 10859 LNCS, pp 15–27. https://doi.org/10.1007/978-3-319-91947-8_2
40. Hercig T, Krejzl P, Hourová B, Steinberger J, Lenc L (2017) Detecting stance in Czech news commentaries. ITAT, pp 176–180
41. Küçük D, Can F (2018) Stance detection on tweets: an svm-based approach. arXiv, 1–13
42. Kochkina E, Liakata M, Augenstein I (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 evaluations (SemEval-2017), pp 475–480
43. Vamvas J, Sennrich R (2020) X-stance: a multilingual multi-target dataset for stance detection. In: 5th SwissText & 16th KONVENS Joint Conference 2020
44. Zotova E, Agerri R, Rigau G (2021) Semi-automatic generation of multilingual datasets for stance detection in twitter. Expert Syst Appl 170:1–29. https://doi.org/10.1016/j.eswa.2020.114547
45. Hutto CJ, Gilbert E (2014) Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the 8th international conference on weblogs and social media, ICWSM 2014
46. Mohammad SM, Turney PD (2013) Crowdsourcing a word-emotion association lexicon. Comput Intell. https://doi.org/10.1111/j.1467-8640.2012.00460.x
47. Pamungkas EW, Basile V, Patti V (2019) Stance classification for rumour analysis in twitter: exploiting affective information and conversation structure. In: 2nd international workshop on rumours and deception in social media (RDSM), pp 1–7
48. Sobhani P, Mohammad SM, Kiritchenko S (2016) Detecting stance in tweets and analyzing its interaction with sentiment. In: Proceedings of the fifth joint conference on lexical and computational semantics (SEM 2016), pp 159–169
49. Zhang B, Yang M, Li X, Ye Y, Xu X, Dai K (2020) Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 3188–3197
50. Mohammad SM, Sobhani P, Kiritchenko S (2017) Stance and sentiment in tweets. ACM Trans Internet Technol (TOIT) 17:1–23
51. Vychegzhanin S, Kotelnikov E (2021) A new method for stance detection based on feature selection techniques and ensembles of classifiers. IEEE Access 9:134899–134915. https://doi.org/10.1109/ACCESS.2021.3116657
52. Hu M, Liu B (2004) Mining and summarizing customer reviews. In: KDD-2004 - proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. https://doi.org/10.1145/1014052.1014073
53. Kobbe J, Hulpus I, Stuckenschmidt H (2020) Unsupervised stance detection for arguments from consequences. Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 50–60
54. Li Y, Caragea C (2019) Multi-task stance detection with sentiment and stance lexicons. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, pp 6299–6305
55. Lai M, Cignarella AT, Farías DIH, Bosco C, Patti V, Rosso P (2020) Multilingual stance detection in social media political debates. Comput Speech Lang 63:1–27. https://doi.org/10.1016/j.csl.2020.101075
56. Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: HLT/EMNLP 2005 - Human Language technology conference and conference on empirical methods in natural language processing, proceedings of the conference. https://doi.org/10.3115/1220575.1220619
57. Dey K, Shrivastava R, Kaushik S (2017) Twitter stance detection-a subjectivity and sentiment polarity inspired two-phase approach. In: IEEE international conference on data mining workshops (ICDMW), pp 365–372
58. Pennebaker JW, Booth RJ, Boyd RL, Francis ME (2001) Linguistic inquiry and word count: Liwc2001. Lawrence Erlbaum Associates 71
59. Ebrahimi J, Dou D, Lowd D (2016) Weakly supervised tweet stance classification by relational bootstrapping. In: proceedings of the 2016 conference on empirical methods in natural language processing, pp 1012–1017
60. Whissell C (2009) Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. Psychol Rep. https://doi.org/10.2466/PR0.105.2.509-521
61. Årup Nielsen F (2011) A new anew: evaluation of a word list for sentiment analysis in microblogs. CEUR Workshop Proceedings, vol. 718
62. Cambria E, Li Y, Xing FZ, Poria S, Kwok K (2020) Senticnet 6: ensemble application of symbolic and subsymbolic ai for sentiment analysis. In: International conference on information and knowledge management. https://doi.org/10.1145/3340531.3412003
63. Liu R, Lin Z, Tan Y, Wang W (2021) Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021:3152–3157
64. Luo Y, Liu Z, Shi Y, Zhang Y (2022) Exploiting sentiment and common sense for zero-shot stance detection. arxiv:2208.08797
65. Speer R, Chin J, Havasi C (2017) Conceptnet 5.5: an open multilingual graph of general knowledge. In: Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), pp 4444–4451
66. Aldayel A, Magdy W (2019) Your stance is exposed! analysing possible factors forstance detection on social media. Proc ACM Hum-Comput Interact 3:1–20
67. Sobhani P, Inkpen D, Matwin S (2015) From argumentation mining to stance classification. In: Proceedings of the 2nd workshop on argumentation mining, pp 67–77
68. Chen W-F, Ku L-W (2016) Utcnn: a deep learning model of stance classificationon on social media text. In: Proceedings of COLING 2016, the 26th International conference on computational linguistics, pp 1635–1645
69. Zarrella G, Marsh A (2016) Mitre at semeval-2016 task 6: transfer learning for stance detection. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016), pp 458–463
70. Ebrahimi J, Dou D, Lowd D (2016) A joint sentiment-target-stance model for stance classification in tweets, pp 2656–2665
71. Wei W, Zhang X, Liu X, Chen W, Wang T (2016) pkudblab at semeval-2016 task 6 : a specific convolutional neural network system for effective stance detection. In: Proceedings of SemEval-2016, pp 384–388
72. Hacohen-Kerner Y, Ido Z, Ya'akobov R (2017) Stance classification of tweets using skip char ngrams. Joint European conference on machine learning and knowledge discovery in databases, pp 266–278
73. Zhou Y, Cristea A, Shi L (2017) Connecting targets to tweets: semantic attention-based model for target-specific stance

detection, pp 18–32. Springer, Cham. https://doi.org/10.1007/978-3-319-68783-4_2
74. Lai M, Cignarella AT, Irazúas H (2017) itacos at ibereval2017: detecting stance in catalan and spanish tweets. In: Proceedings of the second workshop on evaluation of human language technologies for Iberian Languages (IberEval 2017), pp 185–192
75. Du J, Xu R, He Y, Gui L (2017) Stance classification with target-specific neural attention networks. In: 26th International joint conference on artificial intelligence (IJCAI)
76. Siddiqua UA, Chy AN, Aono M (2018) Stance detection on microblog focusing on syntactic tree representation. In: International conference on data mining and big data, vol. 10943, pp 478–490. Springer, Cham. https://doi.org/10.1007/978-3-319-93803-5_45
77. Sun Q, Wang Z, Zhu Q, Zhou G (2018) Stance detection with hierarchical attention network. In: Proceedings of the 27th international conference on computational linguistics, pp 2399–2409
78. Benton A, Dredze M (2018) Using author embeddings to improve tweet stance classification, pp 184–194
79. Wei P, Mao W, Zeng D (2018) A target-guided neural memory model for stance detection in twitter, pp 1–8
80. Sun L, Li X, Zhang B, Ye Y, Xu B (2019) Learning stance classification with recurrent neural capsule network. In: CCF international conference on natural language processing and Chinese computing, pp 277–289
81. Sun Q, Wang Z, Li S, Zhu Q, Zhou G (2019) Stance detection via sentiment information and neural network model. Front Comp Sci 13:127–138. https://doi.org/10.1007/s11704-018-7150-9
82. Lai M, Tambuscio M, Patti V, Ruffo G, Rosso P (2019) Stance polarity in political debates: a diachronic perspective of network homophily and conversations on twitter. Data Knowl Eng. https://doi.org/10.1016/j.datak.2019.101738
83. Li W, Xu Y, Wang G (2019) Stance detection of microblog text based on two-channel cnn-gru fusion network. IEEE Access 7:145944–145952. https://doi.org/10.1109/ACCESS.2019.2944136
84. Wei P, Mao W, Chen G (2019) A topic-aware reinforced model for weakly supervised stance detection, pp 7249–7256
85. Chauhan DS, Kumar R, Ekbal A (2019) Attention based shared representation for multi-task stance detection and sentiment analysis, vol. 1143, pp 661–669. Springer, Cham. https://doi.org/10.1007/978-3-030-36802-9_70
86. Tshimula JM, Chikhaoui B, Wang S (2020) A pre-training approach for stance classification in online forums, pp 280–287
87. Mohtarami M, Glass J, Nakov P (2019) Contrastive language adaptation for cross-lingual stance detection, pp 4442–4452
88. Hosseinia M, Dragut E, Mukherjee A (2019) Pro/con: neural detection of stance in argumentative opinion pro/con: Neural detection of stance in argumentative opinions. In: International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation, pp 21–30
89. Zhou S, Lin J, Tan L, Liu X (2019) Condensed convolution neural network by attention over self-attention for stance detection in twitter, pp 1–8
90. Zhu L, He Y, Zhou D (2020) Neural opinion dynamics model for the prediction of user-level stance dynamics. Inf Process Manage 57:1–13. https://doi.org/10.1016/j.ipm.2019.03.010
91. Al-Ghadir AI, Azmi AM, Hussain A (2021) A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments. Inform Fus 67:29–40. https://doi.org/10.1016/j.inffus.2020.10.003
92. Ahmed M, Chy AN, Chowdhury NK (2020) Incorporating hand-crafted features in a neural network model for stance detection on microblog. In: The 6th international conference on communication and information processing, pp 57–64. Association for Computing Machinery, NY, USA. https://doi.org/10.1145/3442555.3442565
93. Yang Y, Wu B, Zhao K, Guo W (2020) Tweet stance detection: a two-stage dc-bilstm model based on semantic attention. In: IEEE Fifth International conference on data science in cyberspace (DSC), pp 22–29. https://doi.org/10.1109/DSC50466.2020.00012
94. Rashed A, Kutlu M, Darwish K, Elsayed T, Bayrak C (2020) Embeddings-based clustering for target specific stances: The case of a polarized turkey. In: Proceedings of the International AAAI Conference on web and social media, pp 537–548
95. Darwish K, Stefanov P, Aupetit M, Nakov P (2020) Unsupervised user stance detection on twitter. In: Proceedings of the fourteenth international aaai conference on web and social media (ICWSM), pp 141–152
96. Samih Y, Darwish K (2021) A few topical tweets are enough for effective user stance detection, pp 2637–2646
97. Giorgioni S, Politi M, Salman S, Croce D, Basili R (2020) Unitor @ sardistance2020: combining transformer-based architectures and transfer learning for robust stance detection. EVALITA Evaluation of NLP and Speech Tools for Italian
98. Zhao G, Yang P (2020) Pretrained embeddings for stance detection with hierarchical capsule network on social media. ACM Trans Inform Syst 39:1–32. https://doi.org/10.1145/3412362
99. Yang D, Wu Q, Chen W, Wang T, Qiu Z, Liu D, Cui Y (2020) Stance detection with stance-wise convolution network. In: CCF International conference on natural language processing and Chinese computing, vol. 12430 LNAI, pp 555–567. https://doi.org/10.1007/978-3-030-60450-9_44
100. Chen P, Ye K, Cui X (2021) Integrating n-gram features into pre-trained model: a novel ensemble model for multi-target stance detection. In: International conference on artificial neural networks, vol. 12893, pp 269–279. https://doi.org/10.1007/978-3-030-86365-4_22
101. Alkhalifa R, Kochkina E, Zubiaga A (2021) Opinions are made to be changed: temporally adaptive stance classification. In: Proceedings of the 2021 workshop on open challenges in online social networks, pp 27–32. Association for Computing Machinery, NY, USA. https://doi.org/10.1145/3472720.3483620
102. Schiller B, Daxenberger J, Gurevych I (2021) Stance detection benchmark: How robust is your stance detection? KI-Künstliche Intelligenz 35:329–341. https://doi.org/10.1007/s13218-021-00714-w
103. Kawintiranon K, Singh L (2021) Knowledge enhanced masked language model for stance detection. In: Proceedings of the 2021 conference of the North American Chapter of the association for computational linguistics: human language technologies, pp 4725–4735
104. Cotfas LA, Delcea C, Roxin I, Ioanăş C, Gherai DS, Tajariol F (2021) The longest month: analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. IEEE Access 9:33203–33223. https://doi.org/10.1109/ACCESS.2021.3059821
105. Li Y, Zhao C, Caragea C (2021) Improving stance detection with multi-dataset learning and knowledge distillation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp 6332–6345
106. Gómez-Suta M, Echeverry-Correa J, Soto-Mejía JA (2023) Stance detection in tweets: a topic modeling approach supporting explainability. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2022.119046

107. Aker A, Derczynski L, Bontcheva K (2017) Simple open stance classification for rumour analysis. In: Proceedings of the international conference recent advances in natural language processing, RANLP, pp 31–39
108. Bahuleyan H, Vechtomova O (2017) Uwaterloo at semeval-2017 task 8: detecting stance towards rumours with topic independent features. In: Proceedings of the 11th international workshop on semantic evaluations (SemEval-2017), pp 461–464
109. Mohtarami M, Baly R, Glass J, Nakov P, Marquez L, Moschitti A (2018) Automatic stance detection using end-to-end memory networks, pp 767–776
110. Ma J, Gao W, Wong KF (2018) Detect rumor and stance jointly by neural multi-task learning. Companion proceedings of the web conference 2018, pp 585–593. Association for Computing Machinery, NY, USA. https://doi.org/10.1145/3184558.3188729
111. Poddar L, Hsu W, Lee ML, Subramaniyam S (2018) Predicting stances in twitter conversations for detecting veracity of rumors: A neural approach. In: IEEE 30th international conference on tools with artificial intelligence, ICTAI, vol 2018, pp 65–72. https://doi.org/10.1109/ICTAI.2018.00021
112. Bhatt G, Sharma A, Sharma S, Nagpal A, Raman B, Mittal A (2018) Combining neural, statistical and external features for fake news stance identification, pp 1353–1357. Association for Computing Machinery, NY, USA. https://doi.org/10.1145/3184558.3191577
113. Zubiaga A, Kochkina E, Liakata M, Procter R, Lukasik M, Bontcheva K, Cohn T, Augenstein I (2018) Discourse-aware rumour stance classification in social media using sequential classifiers. Inf Process Manage 54:273–290. https://doi.org/10.1016/j.ipm.2017.11.009
114. Borges L, Martins B, Calado P (2019) Combining similarity features and deep representation learning for stance detection in the context of checking fake news. J Data Inform Qual (JDIQ) 11:1–26. https://doi.org/10.1145/3287763
115. Lukasik M, Bontcheva K, Cohn T, Zubiaga A, Liakata M, Procter R (2019) Gaussian processes for rumour stance classification in social media. ACM Trans Inform Syst 37:1–24. https://doi.org/10.1145/3295823
116. Popat K, Mukherjee S, Yates A, Weikum G (2019) Stancy: stance classification based on consistency cues. 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), pp 6413–6418
117. Wei P, Xu N, Mao W (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), pp 4787–4798
118. Yang R, Xie W, Liu C, Yu D (2019) Blcu nlp at semeval-2019 task 7: an inference chain-based gpt model for rumour evaluation, pp 1090–1096
119. Xuan K, Xia R (2019) Rumor stance classification via machine learning with text, user and propagation features. In: IEEE International Conference on Data Mining Workshops, ICDMW, vol 2019, pp 560–566. https://doi.org/10.1109/ICDMW.2019.00085
120. Bugueño M, Mendoza M (2019) Applying self-attention for stance classification. Iberoamerican Congress on Pattern Recognition, pp 51–61
121. Fajcik M, Burget L, Smrz P (2019) But-fit at semeval-2019 task 7: determining the rumour stance with pre-trained deep bidirectional transformers, pp 1097–1104
122. Fang W, Nadeem M, Mohtarami M, Glass J (2019) Neural multi-task learning for stance prediction. In: Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), pp 13–19
123. Islam MR, Muthiah S, Ramakrishnan N (2019) Rumorsleuth: joint detection of rumor veracity and user stance. In: 2019 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp 131–136. Association for Computing Machinery, NY, USA. https://doi.org/10.1145/3341161.3342916
124. Prakash A, Madabushi HT (2020) Incorporating count-based features into pre-trained models for improved stance detection. In: Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propagand, pp 22–32
125. Körner E, Wiedemann G, Hakimi AD, Heyer G, Potthast M (2021) On classifying whether two texts are on the same side of an argument. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 10130–10138. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic
126. Yang S, Urbani J (2021) Tribrid: Stance classification with neural inconsistency detection. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 6831–6843
127. Ye K, Piao Y, Zhao K, Cui X (2021) Graph enhanced bert for stance-aware rumor verification on social media. In: International conference on artificial neural networks, vol. 12895 LNCS, pp 422–435. Springer, Cham. https://doi.org/10.1007/978-3-030-86383-8_34
128. Khandelwal A (2021) Fine-tune longformer for jointly predicting rumor stance and veracity. In: 3rd ACM India Joint international conference on data science and management of data, CODS-COMAD 2021, pp 10–19. Association for Computing Machinery, NY, USA. https://doi.org/10.1145/3430984.3431007
129. Roy A, Fafalios P, Ekbal A, Zhu X, Dietze S (2021) Exploiting stance hierarchies for cost-sensitive stance detection of web documents. J Intelli Inform Syst. https://doi.org/10.1007/s10844-021-00642-z
130. Zhang H, Qian S, Fang Q, Xu C (2021) Multi-modal meta multi-task learning for social media rumor detection. IEEE Trans Multimed. https://doi.org/10.1109/TMM.2021.3065498
131. Augenstein I, Rocktäschel T, Vlachos A, Bontcheva K (2016) Stance detection with bidirectional conditional encoding. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 876–885
132. Xu C, Paris C, Nepal S, Sparks R (2018) Cross-target stance classification with self-attention networks. Proceedings of the 56th annual meeting of the association for computational linguistics, pp 778–783
133. Liang B, Fu Y, Gui L, Yang M, Du J, He Y, Xu R (2021) Target-adaptive graph for cross-target stance detection. In: Proceedings of the world wide web conference, WWW 2021, pp 3453–3464. Association for Computing Machinery, NY, USA. https://doi.org/10.1145/3442381.3449790
134. Hardalov M, Arora A, Nakov P, Augenstein I (2021) Cross-domain label-adaptive stance detection. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 9011–9028
135. Allaway E, Srikanth M, Mckeown K (2021) Adversarial learning for zero-shot stance detection on social media. In: Proceedings of the 2021 conference of the North American Chapter of the association for computational linguistics: human language technologies, pp 4756–4767
136. Conforti C, Berndt J, Pilehvar MT, Giannitsarou C, Toxvaerd F, Collier N (2021) Synthetic examples improve cross-target generalization: a study on stance detection on a twitter corpus. In:

Proceedings of the 11th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp 181–187
137. Sun Q, Xi X, Sun J, Wang Z, Xu H (2022) Stance detection with a multi-target adversarial attention network. ACM Trans Asian Low-Resour Lang Inform Process. https://doi.org/10.1145/3544490
138. Liu Y, Zhang XF, Wegsman D, Beauchamp N, Wang L (2022) Politics: pretraining with same-story article comparison for ideology prediction and stance detection, pp 1354–1374. arxiv:2205.00619
139. Fu Y, Li X, Li Y, Wang S, Li D, Liao J, Zheng J (2022) Incorporate opinion-towards for stance detection. Knowl-Based Syst 246:1–11. https://doi.org/10.1016/j.knosys.2022.108657
140. Wei P, Lin J, Mao W (2018) Multi-target stance detection via a dynamic memory-augmented network. In: The 41st international ACM SIGIR conference on research & development in information retrieval, pp 1229–1232. Association for Computing Machinery, NY, USA. https://doi.org/10.1145/3209978.3210145
141. Sobhani P, Inkpen D, Zhu X (2019) Exploring deep neural networks for multitarget stance detection. Comput Intell 35:82–97. https://doi.org/10.1111/coin.12189
142. Siddiqua UA, Chy AN, Aono M (2019) Tweet stance detection using multi-kernel convolution and attentive lstm variants. IEICE Trans Inf Syst 102:2493–2503. https://doi.org/10.1587/transinf.2019EDP7080
143. Li Y, Caragea C (2021) A multi-task learning framework for multi-target stance detection, pp 2320–2326
144. Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor Newslett 10(1145/3137597):3137600
145. Vashishth S, Sanyal S, Nitin V, Talukdar PP (2020) Composition-based multirelational graph convolutional networks. In: 8th international conference on learning representations, ICLR 2020, Addis Ababa, Ethiopia, Apr 26- 30
146. El-Alfy E-SM, Luqman H (2022) A comprehensive survey and taxonomy of sign language research. Eng Appl Artif Intell 114:105198
147. Sagi O, Rokach L (2018) Ensemble learning: a survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. https://doi.org/10.1002/widm.1249
148. Vychegzhanin SV, Kotelnikov EV (2019) Stance detection based on ensembles of classifiers. Prog Comput Softw 45:228–240. https://doi.org/10.1134/S0361768819050074
149. Vychegzhanin S, Razova E, Kotelnikov E, Milov V (2019) Selecting an optimal feature set for stance detection. In: International Conference on analysis of images, social networks and texts, vol. 11832 LNCS. https://doi.org/10.1007/978-3-030-37334-4_22
150. Vychegzhanin SV, Razova EV, Kotelnikov EV (2019) What number of features is optimal? a new method based on approximation function for stance detection task. In:Proceedings of the 9th international conference on information communication and management, pp 43–47. https://doi.org/10.1145/3357419.3357430
151. Ruder S (2019) Neural transfer learning for natural language processing. PhD thesis, National University of Ireland, Galway. http://hdl.handle.net/10379/15463
152. Margolis A (2011) A literature review of domain adaptation with unlabeled data. Tec, Report
153. Alec R, Jeffrey W, Rewon C, David L, Dario A, Ilya S (2019) Language models are unsupervised multitask learners. OpenAI Blog 1
154. Devlin J, Chang MW, Lee K, Toutanova K (2019) Bert: pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019 - 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1
155. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. https://doi.org/10.18653/v1/n18-1202
156. Clark K, Luong MT, Le QV, Manning CD (2020) Electra: pre-training text encoders as discriminators rather than generators. arXiv
157. Ruder S, Peters M, Swayamdipta S, Wolf T (2019) Transfer learning in natural language processing tutorial. In: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Tutorial Abstracts
158. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
159. Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) 1, pp 328–339. https://doi.org/10.18653/v1/p18-1031
160. Ruder S (2017) An overview of multi-task learning in deep neural networks. arXiv
161. Wang Y, Yao Q, Kwok JT, Ni LM (2020) Generalizing from a few examples: a survey on few-shot learning. ACM Comput Surv. https://doi.org/10.1145/3386252
162. Wang Y, Yao Q, Kwok JT, Ni LM (2020) Generalizing from a few examples: a survey on few-shot learning. ACM Comput Surv (csur) 53(3):1–34
163. Zhang Y, Yang Q (2017) A survey on multi-task learning. arXiv
164. Tsarfaty R, Bareket D, Klein S, Seker A (2020) From spmrl to nmrl: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (mrls)? arXiv. https://doi.org/10.18653/v1/2020.acl-main.660
165. Ghosh S, Singhania P, Singh S, Rudra K, Ghosh S (2019) Stance detection in web and social media: A comparative study. In: International conference of the cross-language evaluation forum for European Languages, pp 75–87. https://doi.org/10.1007/978-3-030-28577-7_4
166. Aldayel A, Magdy W (2019) Assessing sentiment of the expressed stance on social media. In: International Conference on Social Informatics, pp 277–286. https://doi.org/10.1007/978-3-030-34971-4_19
167. Li Y, Tian X, Liu T, Tao D (2015) Multi-task model and feature joint learning. IJCAI International Joint Conference on Artificial Intelligence, pp 3643–3649
168. Kitchenham B (2007) Guidelines for performing systematic literature reviews in software engineering. Technical report, Ver. 2.3 EBSE Technical Report. EBSE