Stance Detection: A Survey

DILEK KÜÇÜK, TÜBİTAK Energy Institute
FAZLI CAN, Bilkent University

Automatic elicitation of semantic information from natural language texts is an important research problem with many practical application areas. Especially after the recent proliferation of online content through channels such as social media sites, news portals, and forums; solutions to problems such as sentiment analysis, sarcasm/controversy/veracity/rumour/fake news detection, and argument mining gained increasing impact and significance, revealed with large volumes of related scientific publications. In this paper, we tackle an important problem from the same family and present a survey of *stance detection* in social media posts and (online) regular texts. Although stance detection is defined in different ways in different application settings, the most common definition is "automatic classification of the stance of the producer of a piece of text, towards a target, into one of these three classes: {*Favor*, *Against*, *Neither*}." Our survey includes definitions of related problems and concepts, classifications of the proposed approaches so far, descriptions of the relevant datasets and tools, and related outstanding issues. Stance detection is a recent natural language processing topic with diverse application areas, and our survey paper on this newly-emerging topic will act as a significant resource for interested researchers and practitioners.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Machine learning**; **Language resources**; • **Information systems** → **Web and social media search**; **Sentiment analysis**.

Additional Key Words and Phrases: Stance detection, Twitter, Social media analysis, Deep learning

ACM Reference Format:
Dilek Küçük and Fazli Can. 2019. Stance Detection: A Survey. *ACM Comput. Surv.* 1, 1, Article 1 (January 2019), 37 pages. https://doi.org/10.1145/3369026

1 INTRODUCTION

Automatic information extraction from texts is an important research topic of natural language processing (NLP) for decades. Recent widespread use of online and publicly-available tools leads to the accumulation of large volumes of textual content ready to be analyzed for various practical purposes. These tools include news portals, user forums, blogs, publishing platforms, and social media sites like Twitter, Facebook, and Instagram. Some of the main research problems regarding the automatic analysis of this content include sentiment analysis (opinion mining), emotion recognition, argument mining (reason identification), sarcasm/irony detection, veracity and rumour detection, and fake news detection. Automatic and high-performance solutions to these problems will facilitate important tasks ranging from trend and market analysis, obtaining user reviews for products, opinion surveys, targeted advertising, polling,

Authors' addresses: Dilek Küçük, TÜBİTAK Energy Institute, Electrical Power Technologies Department, dilek.kucuk@tubitak.gov.tr; Fazli Can, Bilkent University, Bilkent Information Retrieval Group, Computer Engineering Department, canf@cs.bilkent.edu.tr.

predictions for elections and referendums, automatic media monitoring, and filtering out unconfirmed content for better user experience, to online public health surveillance.

Stance detection (also known as stance classification [Walker et al. 2012a], stance identification [Zhang et al. 2017], stance prediction [Qiu et al. 2015], debate-side classification [Anand et al. 2011], and debate stance classification [Hasan and Ng 2013]) is a considerably recent member of the aforementioned family of research problems. It is usually considered as a subproblem of sentiment analysis and aims to identify the stance of the text author towards a target (an entity, concept, event, idea, opinion, claim, topic, etc.) either explicitly mentioned or implied within the text [Mohammad et al. 2016b; Sobhani 2017]. Although they evolve around this basic purpose and hence are semantically close, there are three mainstream definitions regarding the stance detection problem (some in distinct problem settings) as reported in the literature, namely, generic *stance detection* [Mohammad et al. 2016b], *rumour stance classification* [Zubiaga et al. 2018a], and *fake news stance detection* [FNC 2017]. Based on the number of targets, and the existence of the stance target in the training and testing datasets of the experimental settings, two other subclasses of the initial generic stance detection problem can be defined: *multi-target stance detection* [Sobhani 2017] and *cross-target stance detection* [Augenstein et al. 2016a; Xu et al. 2018].

Prior to presenting these definitions, it will be useful to provide a definition of stance itself from a point of view in linguistics. Hence, Du Bois describes stance as follows: "*Stance is a public act by a social actor, achieved dialogically through overt communicative means, of simultaneously evaluating objects, positioning subjects (self and others), and aligning with other subjects, with respect to any salient dimension of the sociocultural field*" [Du Bois 2007]. Hence, based on this definition, in a stance act, a stancetaker reveals its evaluation on an object and thereby aligns herself/himself with others [Du Bois 2007]. Interested readers are referred to [Du Bois 2007] for further details on a linguistics-based unified framework of stance.

Returning back to the process of automatic stance detection, aforementioned definitions of stance detection are provided below.

Definition 1.1 (Stance Detection). For an input in the form of a piece of text and a target pair, stance detection is a classification problem where the stance of the author of the text is sought in the form of a category label from this set: {*Favor*, *Against*, *Neither*}. Occasionally, the category label of *Neutral* is also added to the set of stance categories [Mohammad et al. 2016b] and the target may or may not be explicitly mentioned in the text [Augenstein et al. 2016a; Mohammad et al. 2016b].

Definition 1.2 (Multi-target Stance Detection). For an input in the form of a piece of text and a set of related targets, multi-target stance detection is a classification problem where the stance of the text author is sought as a category label from this set: {*Favor*, *Against*, *Neither*} for each target and each stance classification (for each target) might have an effect on the classifications for the remaining targets [Sobhani 2017].

Definition 1.3 (Cross-target Stance Detection). Cross-target stance detection is a classification problem where the stance of the text author is sought for a specific target as a category label from this set: {*Favor*, *Against*, *Neither*}, in a settings where stance annotations are available for (though related but) different targets, i.e., there is not enough stance-annotated training data for the target under consideration [Augenstein et al. 2016a; Xu et al. 2018].

Definition 1.4 (Rumour Stance Classification). For an input in the form of a piece of text and a rumour pair, rumour stance classification is a problem where the position of the text author towards the veracity of the rumour is sought

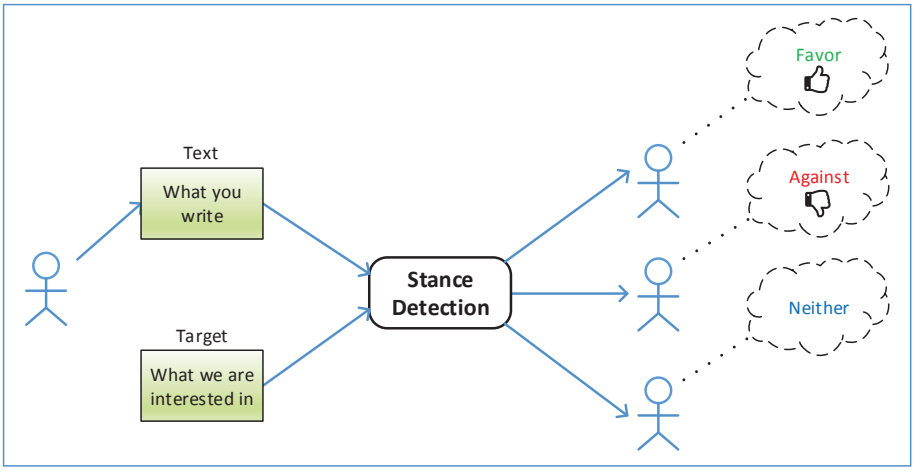


Fig. 1. Schematic representation of the stance detection procedure.

for, in the form of a category label from this set: {*Supporting*, *Denying*, *Querying*, *Commenting*}. As the set of possible category labels, a subset of this set such as {*Supporting*, *Denying*} is occasionally employed [Zubiaga et al. 2018a].

Definition 1.5 (Fake News Stance Detection). For an input in the form of news headline and a news body pair (where the headline and body parts may belong to different news articles), this is a classification problem where the stance of the body towards the claim of the headline is sought for, in the form of a category label from this set: {*Agrees*, *Disagrees*, *Discusses* (the same topic), *Unrelated*}. This problem is defined in order to facilitate the task of fake news detection [FNC 2017].

The most common definition of automatic stance detection, as observed in the related literature, is the first one given above. If we state this definition in other words, stance detection is predicting one's stance on what we are interested in what she/he writes. This definition is depicted schematically in Figure 1.

In this paper, we present a comprehensive survey of automatic stance detection in regular texts and social media posts. Stance detection is an NLP problem still in its nascent stage, yet, there is a considerable body of conducted research on the topic. Hence, a plausible review of the related literature on stance detection will hopefully stand as an important contribution to the topics of social media analysis, NLP, and machine learning. In other words, this paper addresses the need for a comprehensive survey on the recent and significant research topic of stance detection, by putting it into perspective with respect to the related problems and by presenting in-depth information on its historical evolution, classification approaches to the problem, its related datasets, application areas, and open research issues. Hence, this survey paper will help researchers and practitioners of stance detection identify the main approaches, feature sets, best practices, and open issues to start conducting on-topic research and building automatic systems, in addition to the related software tools and datasets that will facilitate related research and development efforts.

The rest of this section includes information on the organization and content of the remaining sections of the paper. As mentioned earlier, stance detection is related to a number of important research problems in NLP. These problems and the corresponding interrelationships are elaborated in Section 2. A generic and common system architecture (with shallow differences seen in different studies) for stance detection is described in Section 3. Earlier work on

stance detection which are conducted mostly on online debate posts using traditional classification algorithms, and stance detection competitions performed so far are described in Section 4. Traditional feature-based machine learning algorithms are commonly used both in earlier work as well as in recent work on stance detection. More recent studies also apply deep learning techniques and ensemble algorithms combining several classifiers. Based on this categorization, approaches to stance detection are reviewed in Section 5. Although limited in number and diversity, there are annotated stance datasets, annotation guidelines, and evaluation metrics used for stance detection, as reported in the related studies. These resources and metrics are described in Section 6. Software and other tools built or used for stance detection purposes are presented in Section 7. Stance detection is known to have several practical application areas such as polling, trend analysis, automatic summarization, and rumour or fake news detection. These application areas constitute the focus of Section 8. Being a research problem in its earlier years, there are several outstanding issues regarding stance detection that need further and considerable research attention. Pointers to such issues are provided in Section 9, and finally Section 10 concludes the paper with a summary. Our paper also has online *supplementary metarial* that includes “some remarks on approaches to stance detection" and “observations and recommendations for stance detection researchers". They respectively extend Section 5 and Section 9, and can be accessed via the link provided in the ACM Digital Library.

2 STANCE DETECTION AND RELATED PROBLEMS

As given in the definitions stated in the previous section, stance detection in natural language texts is concerned with the position (or stance) of the text producer towards a target or a set of targets. Initial studies on stance detection aim to determine the stance of the people in online debate forums towards ideological or controversial issues. More recently, a stance detection competition on tweets is performed in 2016 within the course of the annual *Workshop on Semantic Evaluation (SemEval)*[1], based on Definition 1.1 of Section 1, and a stance detection competition for fake news detection is established in 2017 (named *Fake News Challenge*)[2], based on Definition 1.5. Stance detection is also employed in rumour detection pipelines, based on Definition 1.4 of Section 1. With this progress, the domain text genres of stance detection now commonly includes social media texts, news articles, online user comments on news, as well.

Table 1. Sample Tweets from SemEval 2016 Stance Dataset [Mohammad et al. 2016b].

Tweet	**Stance Target**	**Stance**	**Sentiment**
RT @TheCLF: Thanks to everyone in Maine who contacted their legislators in support of #energyefficiency funding! #MEpoli #SemST	Climate Change is a Real Concern	Favor	Positive
We live in a sad world when wanting equality makes you a troll... #SemST	Feminist Movement	Favor	Negative
I don’t believe in the hereafter. I believe in the here and now. #SemST	Atheism	Favor	Neither
@violencehurts @WomenCanSee The unborn also have rights #defendthe8th #SemST	Legalization of Abortion	Against	Positive
I’m conservative but I must admit I’d rather see @SenSanders as president than Mrs. Clinton. #stillvotingGOP #politics #SemST	Hillary Clinton	Against	Negative

[1] http://alt.qcri.org/semeval2016/task6/
[2] http://www.fakenewschallenge.org/

I have my work and my faith... If that's boring to some people, I can't tell you how much I don't care. ~Madonna Ciccone #SemST	Atheism	Against	Neither
@BadgerGeno @kreichert27 @jackbahlman Too busy protesting :) #LoveForAll #BackdoorBadgers #SemST	Hillary Clinton	Neither	Positive
@ShowTruth You're truly unwelcome here. Please leave. #ygk #SemST	Legalization of Abortion	Neither	Negative
@Maisie_Williams everyone feels that way at times. Not just women #SemST	Atheism	Neither	Neither

Table 1 provides sample tweets and stance targets with the corresponding stance and sentiment classifications, from the SemEval 2016 stance dataset [Mohammad et al. 2016b] which was also annotated with sentiment information within the course of a subsequent study [Mohammad et al. 2017]. The set of stance classes in this dataset is {*Favor*, *Against*, *Neither*} and the sentiment classes are from the set: {*Positive*, *Negative*, *Neither*}. Hence, the samples in Table 1 are representatives of all nine classification combinations.

Another stance class encountered in the literature is *Neutral* and is considered different from the *Neither* class. That is, if the stance of a piece of text towards a target is not *Favor* or *Against*, then the author's stance may not be necessarily *Neutral*, but instead no stance information can be extracted from the text alone, hence the appropriate stance class for such texts would be *Neither* [Sobhani 2017]. Hence, *Neither* (or, *None*) stance class usually corresponds to all cases other than *Favor* or *Against* classifications.

As a side note, the domain of stance detection research is mostly textual content, and therefore, this review paper covers related stance detection work mostly on texts. Yet, other media content such as speech (as in [Levow et al. 2014]), image, and video (such as movies and news videos) offers significant and practical opportunities for stance detection and this point is revisited as a line of future work in Section 9.

The keyword "position" (or "stance") in the problem definition evokes other keywords such as sentiments, emotions, opinions and hence reveals its close relationship with a number of other NLP or text mining problems: (1) sentiment analysis, (2) emotion recognition, (3) perspective identification, (4) sarcasm/irony detection, (5) controversy detection, (6) argument mining, and (7) biased language detection. The first two of them are related to the more general topic of *affective computing* [Picard 1997] which deals with automatic analysis of all human affects including sentiments and emotions. A schematic representation of these problems related to stance detection, also covering its different subproblems (defined in Section 1) is presented in Figure 2. The details of these related problems are provided in the rest of this section.

2.1 Stance Detection vs. Sentiment Analysis

Sentiment analysis (or opinion mining) is usually defined as the computational treatment of sentiments and opinions in texts [Liu 2010; Pang and Lee 2008; Ravi and Ravi 2015]. Yet, currently the problem is considered mostly equivalent to the detection of the sentiment polarity of a text producer and hence a classification output usually as *Positive*, *Negative*, or *Neutral* is expected from the sentiment analysis procedure. Regardless of the expected output of the generic problem of sentiment analysis (be it the sentiment, polarity, opinion, or subjectivity), main factors that differentiate sentiment analysis and stance detection problems are that (1) the former problem is concerned with the sentiment without a particular target which is expected by the latter one, and that (2) the sentiment and stance (for a target) within the

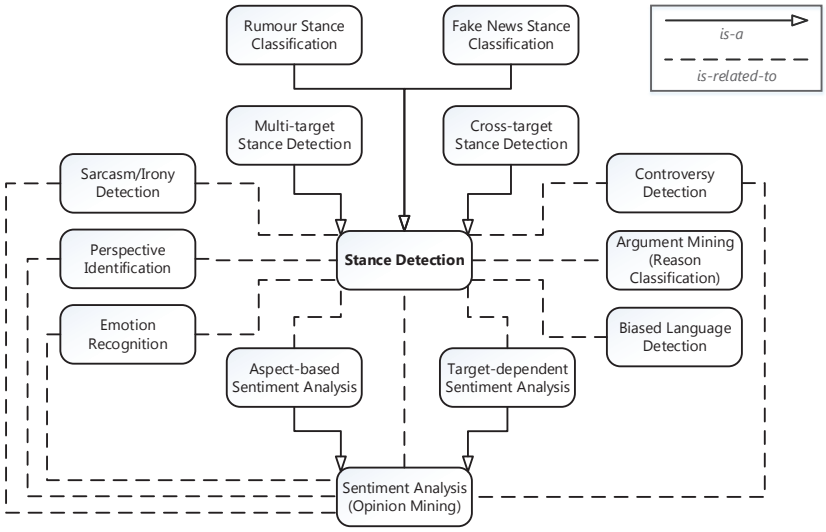


Fig. 2. Research problems related to stance detection and subproblems of stance detection.

same text may not be aligned at all, that is the polarity of the text may be positive while the stance may be against a particular target, and vice versa.

There are two subproblems of sentiment analysis which can be considered more close to stance detection than the generic sentiment analysis problem itself:

(1) *Aspect-oriented (or aspect-based, or aspect-level) sentiment analysis*: In this subproblem of sentiment analysis, the sentiment polarities towards a target entity and different aspects of this entity are considered in a given text input [Pontiki et al. 2015; Schouten and Frasincar 2016]. It is usually considered as a slot-filling problem where three slots are involved: the target entity, the aspect of the entity, the sentiment polarity towards the aspect. In shared datasets for aspect-oriented sentiment analysis, target entities commonly include electronic equipment like laptops, restaurants, and hotels while the corresponding aspects of these entities include price, design, and quality, among others.

(2) *Target-dependent (or target-based) sentiment analysis*: In this subproblem of sentiment analysis, the sentiment polarity towards the target is explored within the text, given a text and target pair [Jiang et al. 2011]. A similar subproblem is *open-domain targeted sentiment analysis* [Mitchell et al. 2013] where both a named entity and the sentiment towards this entity is explored in the input text. As pointed out in [Ebrahimi et al. 2016a], the main differences between stance detection and target-dependent sentiment analysis are: (1) the stance target may not be explicitly given in the input text, (2) the stance target may not be the target of the sentiment in the text. An additional difference is that (3) the stance target may be an event while the target is usually an entity or an aspect in sentiment analysis. These differences also apply to stance detection and open-domain targeted sentiment analysis.

2.2 Stance Detection vs. Emotion Recognition

Emotion recognition (also called emotion detection or emotion extraction) is another task related to stance detection and more closely to sentiment analysis, which aims to extract the emotion from a given text. Emotion recognition can be carried out using limited to more diverse emotion classes. Common emotion classes include *Joy*, *Sadness*, *Anger*, *Disgust*, *Anxiety*, *Surprise*, *Fear*, and *Love*, among others. In various studies on emotion recognition, emotion classes at finer granularity levels than the ones listed here are employed as well. To illustrate, the emotion annotation for the tweet in the first row of Table 1 could possibly be *Joy* while the emotion for the tweet in the second row could be *Sadness*. Interested readers are referred to [Sailunaz et al. 2018] for a survey of emotion recognition studies, and to [Mohammad and Turney 2013] where a word emotion lexicon created through crowdsourcing techniques is presented.

2.3 Stance Detection vs. Perspective Identification

Perspective identification is usually defined as the automatic determination of the point-of-view of the author of a piece of text from its content (such as from the perspective of *Democrats* or *Republicans* in the context of US elections) [Lin et al. 2006; Sobhani 2017; Wong et al. 2016]. Similar to stance detection, it is also related to the subjective evaluation of the text author and hence similarly considered as a topic close to sentiment analysis.

One significant difference between stance detection and perspective identification is that there is a stance target on which the position of the author (usually as *For* or *Against*) is investigated in the former problem while the perspective of the text author from a number of different alternatives (like *Democrats* and *Republicans* for instance) is searched for, without an explicit single topic (or topic group) of consideration, in the latter problem. Yet, as in the case of related research on stance detection, common feature-based machine learning algorithms together with lexical features are also utilized and proved to be useful for the problem of perspective identification [Lin et al. 2006; Wong et al. 2016].

2.4 Stance Detection vs. Sarcasm/Irony Detection

Sarcasm and irony are quite close linguistic phenomena and commonly used interchangeably. In an instance of sarcasm/irony in a piece of text, the text producer utters something different than what s/he actually intends, usually for the purposes of criticism or ridicule. In studies that differentiate the two, sarcasm is defined as the verbal form of an irony.

Sarcasm/irony detection is a classification problem where the existence of sarcasm/irony in a given text is sought for. The problem is considered particularly important for sentiment analysis, as high-performance sarcasm/irony detection in a given text will also improve the performance of the subsequent sentiment analysis procedure, by reverting the sentiment classification output in case of sarcasm/irony detection. More information can be found in [Joshi et al. 2017] and in [Wallace 2015] where surveys of studies on sarcasm detection and irony detection are presented, respectively.

2.5 Stance Detection vs. Controversy Detection

A controversy is usually defined as a discussion regarding a specific target entity which provoke opposing opinions among people, for a finite duration of time [Al-Ayyoub et al. 2018; Popescu and Pennacchiotti 2010]. In controversy detection, a (relevant) controversy score is generally calculated and associated with each unit of content and so that sorting based on those scores can be achieved. The controversy detection problem is also considered very close to the problem of sentiment analysis. In addition to the aforementioned related studies, interested readers are referred to [Dori-Hacohen 2015; Jang and Allan 2016; Jang et al. 2016; Timmermans et al. 2017] for computational treatment of

controversy, which can further be tracked down to Leibniz's idea of *Characteristica Universalis* (or, *Universal Mathematics*) [Russell 1992], or his dream of using calculation for all human reasoning [Dijkstra 1997].

Stance detection is usually performed on controversial topics like debates or elections/referendums. The topic of controversy detection is related to stance detection in the sense that a system for controversy detection can be used as a prospective preprocessing unit for an open-domain stance detection system. That is, currently, stance detection is performed on predefined topics with predefined stance targets, which are usually selected from a set of controversial topics, and controversy detection and stance detection can be performed sequentially (in this order) within a larger automatic system for information elicitation. The former module will help detect the controversial content regarding specific targets and the latter module will help reveal the stances of the content producers towards these targets. The authors in [Zhang et al. 2017] implemented this scheme for public health surveillance by first identifying controversial discussions in online health forums and next detecting stance in the included posts (see Section 8). Hence, studying controversy detection will definitely help researchers and practitioners build similar practical systems in which stance detection module will utilize the output of the controversy detection module.

2.6 Stance Detection vs. Argument Mining

Computational argument (or, argumentation) mining is a recent topic in NLP and deals with the extraction of possible argument structure in a given textual content [Lippi and Torroni 2016]. The main stages of a generic argument mining system are: (1) detection of the argumentative sentences in the text, (2) extraction of argument components (such as claims and evidences/premises), and (3) forming the final argument graph by connecting the extracted components.

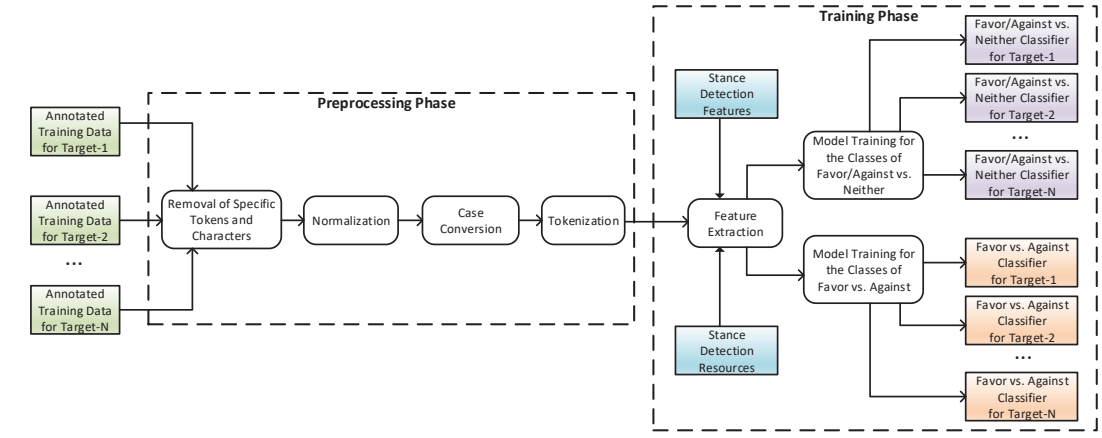
Argument mining is another research topic related to stance detection in the sense that solutions to both of them facilitate automatic understanding of debates/discussions revealed in textual content and related user modeling. Another interrelationship between stance detection and argument mining is that the outputs of argument information can be used to improve the stance detection procedure [Sobhani et al. 2015], or, stance labels can be used within the argument mining procedure [Wojatzki and Zesch 2016b].

2.7 Stance Detection vs. Biased Language Detection

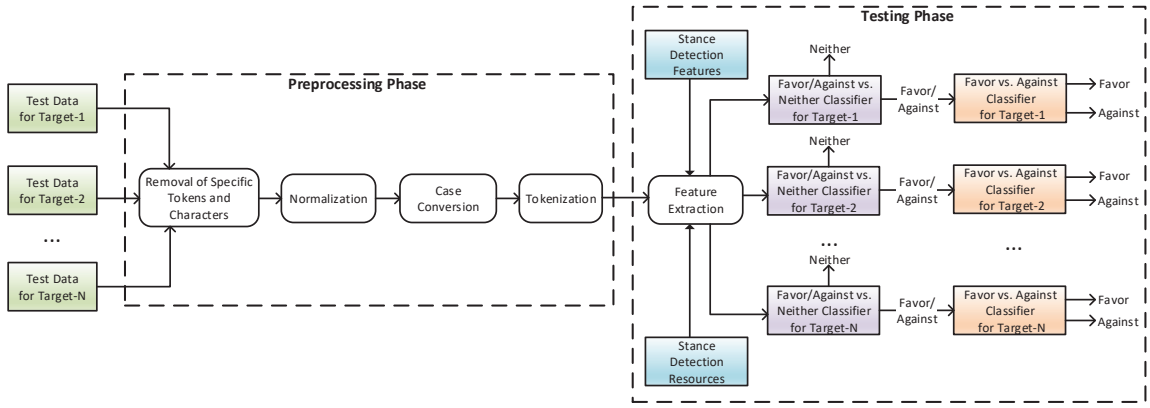
Another research problem closely related to stance detection is biased language detection where the existence of an inclination or tendency towards a particular perspective within a text is explored [Recasens et al. 2013; Yano et al. 2010]. Biased language detection can also be defined as the detection of textual content which includes a particular non-neutral stance. Therefore, based on this definition, a stance detection pipeline may include biased language detection as a subtask. Biased language detection and analysis is particularly useful for online encyclopedias, such as Wikipedia, which are expected to contain information that is free of bias [Recasens et al. 2013].

3 A GENERIC SYSTEM ARCHITECTURE

Stance detection approaches presented in the related literature are learning-based systems including training and testing phases where both of them are accompanied with a preprocessing phase, as commonly observed in recent applied research on different NLP problems. These learning-based approaches can be classified as traditional machine learning, deep learning, and ensemble learning approaches as will be reviewed in Section 5. In this section, we provide a generic system architecture which reflects the common properties of related proposals in the literature. The training and testing phases of this architecture are presented in Figure 3(a) and 3(b).



(a) Training phase.



(b) Testing phase.

Fig. 3. The architecture of a generic stance detection system.

The preprocessing phase is shared and carried out before the actual training and testing phases. The most common tasks performed during the preprocessing phase are:

- *Removal of specific tokens and characters*: Specific tokens like stopwords, URLs, tokens matching the @username pattern (in Twitter mentions and replies[3]), and punctuation marks are removed.
- *Normalization*: In case of tweets, misspelled and contracted forms are normalized.
- *Case conversion*: Tokens are converted to all uppercase or to all lowercase.
- *Tokenization*: This is the last phase before the actual feature selection process of the training/testing phases. The remaining text is split into its individual tokens based on the tokenization rules of the language under consideration.

[3]https://help.twitter.com/en/using-twitter/mentions-and-replies

The corresponding modules in NLP tools such as TweetNLP[4], Stanford CoreNLP[5], and NLTK[6] are commonly used for preprocessing purposes in related studies, as well as proprietary implementations.

During the training phase, stance detection features and resources are utilized to train the classifiers (models). Below are the common characteristics of the training phase of a stance detection system, as reported in the literature:

- For traditional feature-based learning systems (such as Support Vector Machine (SVM) and decision trees), predefined features (such as character and word ngrams, along with features based on POS tags, hashtags, and sentiment dictionaries) are used to train the classifiers. In deep learning approaches, mostly word embedding vectors such as word2vec [Mikolov et al. 2013] trained on large corpora are used as features.
- Training separate classifiers for each stance target is a recommended practice in several relevant studies. Hence, in our generic system architecture for stance detection, we include separate classifiers for each of the targets which are trained individually. An exceptional case to this preference is observed in studies on multi-target stance detection, where for a given piece of input text, a stance class towards multiple targets is expected [Sobhani et al. 2017] (See Definition 1.2 of Section 1). In this case, a single stance classifier for each predefined group of targets is employed instead [Sobhani et al. 2017] due to possible dependencies.
- Again, in many studies it is reported that a pipelined two-phase classification scheme is adequate for three-way stance classification: in the first phase, a classifier determines relevancy, i.e., the input is classified as having a stance (*Favor* or *Against* class) or not (*Neither* class); while in the second phase, the input text classified as having a stance (in the first phase) is further classified as *Favor* or *Against* towards the stance target. Our stance detection architecture also aligns with this scheme, and hence, there are two classifiers trained for each target.

In the testing phase, similarly, preprocessing stages are performed on the input test dataset, and next, for each stance target, two classifiers are applied on the input text in a pipelined manner, in order to output stance as *Favor*, *Against*, or *Neither*.

Our description of the generic architecture so far applies to learning approaches based on a selected single classification algorithm. In order to propose an architecture for ensemble learning approaches to stance detection (see Section 5.1), the components of the proposed architecture should be replicated as needed for each individual classifier considered and a new module implementing the combination algorithm, such as a stacking algorithm, to arrive at the ultimate ensemble classifier is required (see [Bonab and Can 2018]). This final insight concludes our description of a generic architecture for stance detection.

4 A HISTORICAL PERSPECTIVE

In this section, we first review the earlier work on stance detection, considering their particular characteristics. Next, summaries of high-impact stance detection competitions carried out in 2016 and 2017 are provided, where these competitions boosted related research by providing shared annotated datasets, evaluation metrics, and baseline systems.

4.1 Earlier Work on Stance Detection

Distinctive characteristics of the initial studies on stance detection lie in (1) the text genre and annotation characteristics of the datasets that they use and (2) the types of stance detection classifiers and features used by these classifiers.

[4] http://www.cs.cmu.edu/~ark/TweetNLP/
[5] https://stanfordnlp.github.io/CoreNLP/
[6] https://www.nltk.org/

According to the categorization given in [Hasan and Ng 2013], (as of 2013), stance detection studies on debates are mostly conducted on (1) congressional-floor debates [Thomas et al. 2006], (2) company-internal discussions [Murakami and Raymond 2010], and (3) online social, political, and ideological debates [Anand et al. 2011; Somasundaran and Wiebe 2010; Walker et al. 2012a]. Online debates about products [Somasundaran and Wiebe 2009], not mentioned in [Hasan and Ng 2013], can also be added to this list of genres up until 2013. Considering other earlier studies performed after 2013, stance detection experiments on spontaneous speech [Levow et al. 2014], on student essays [Faulkner 2014], and on tweets [Rajadesingan and Liu 2014] are also published. As will be clarified throughout this survey paper, the number of studies on tweets has increased drastically since then, boosted by the related stance detection competitions as overviewed in Section 4.2. Yet, though not comparable in their frequencies with that of the studies on tweets, related studies on online ideological and social debates [Sridhar et al. 2015, 2014] still constitute an important part of the related research on stance detection.

Most of the debate data are obtained from public forums such as `http://convinceme.net` [Anand et al. 2011; Walker et al. 2012a,b], `http://4forums.com` [Misra and Walker 2013], and `http://www.createdebate.com/` [Hasan and Ng 2013]. Common stance targets in online debates include diverse topics including evolution, gun rights, gay rights, abortion, healthcare, death penalty, and existence of God. Table 6 of Section 6.2 includes the stance targets in several available stance detection datasets. Earlier work also demonstrates a slight diversity in the class names used for stance annotation, i.e., in place of the stance classes of {*Favor*, *Against*}, different studies use {*Support*, *Oppose*}, {*Pro*, *Con*}, and {*Pro*, *Anti*}, among others.

In earlier work on stance detection (as well as in recent related work), it is a common practice to employ various different classifiers and compare their performance rates. The classifiers tested in earlier work include rule-based algorithms (such as JRip) [Anand et al. 2011; Murakami and Raymond 2010; Walker et al. 2012a,b]; supervised algorithms like SVM [Hasan and Ng 2013; Somasundaran and Wiebe 2010; Thomas et al. 2006; Walker et al. 2012b], naïve Bayes [Anand et al. 2011; Hasan and Ng 2013; Rajadesingan and Liu 2014; Walker et al. 2012b], boosting [Levow et al. 2014], decision tree and random forest [Misra and Walker 2013], Hidden Markov Models (HMM) and Conditional Random Fields (CRF) [Hasan and Ng 2013]; graph algorithms such as MaxCut [Murakami and Raymond 2010; Walker et al. 2012a], and other approaches such as Integer Linear Programming (ILP) [Somasundaran and Wiebe 2009] and Probabilistic Soft Logic (PSL) [Sridhar et al. 2015, 2014].

One of the distinctive characteristics of the earlier work is that several studies use inter-post information such as agreement/disagreement links, reply links, rebuttal information, and retweeting behavior as important features and it is pointed out in these studies that using such collective information improves stance detection performance compared to processing each post individually. Other common features utilized include word ngrams, cue/topic words, dependencies, argument-related and sentiment/subjectivity features, and frame-semantic features.

Those earlier studies on stance detection which present new stance-annotated datasets are revisited in Section 6.2 (and in the accompanied Table 6), where detailed information about the employed datasets can be found.

4.2 Stance Detection Competitions

To the best of our knowledge, there are three competitions performed on stance detection so far, which are significant as they help boost research on stance detection by providing annotation guidelines, annotated datasets, evaluation metrics, and descriptions of the participating works. These three competitions are (1) SemEval-2016 shared task on stance detection in English tweets, (2) NLPCC-ICCPOL-2016 shared task on stance detection in Chinese microblogs,

and (3) IberEval-2017 shared task on stance detection in Spanish and Catalan tweets. Their details are provided in the following subsections.

4.2.1 SemEval-2016 Task 6: Detecting Stance in Tweets. The earliest competition on stance detection is SemEval-2016 shared task on Twitter stance detection, as described in [Mohammad et al. 2016b]. The competition has two subtasks: in subtask A (*supervised stance detection*), an annotated training dataset of 2,814 tweets and a test dataset of 1,249 tweets are provided for a total of five targets, while in subtask B (*weakly supervised stance detection*), only a large unlabeled dataset (of approximately 78,000 tweets) and a smaller test data (of 707 tweets) for another target are provided to the participants for training and testing, respectively, without any annotated training data. The details of this stance-annotated dataset are provided in [Mohammad et al. 2016a] and also described in Table 6 of Section 6.2.

The participants of the competition employ traditional feature-based machine learning, deep learning, and combined (ensemble) methods. The best performing system for subtask A is a Recurrent Neural Network (RNN) based system [Zarrella and Marsh 2016] and attains an F-score of 67.82% (among all 19 participants of subtask A), while the best system for subtask B is a system based on Convolutional Neural Networks (CNN) achieving an F-score 56.28% (among all 9 participants of subtask B) which also is ranked second for subtask A with an F-score of 67.33% [Wei et al. 2016]. It should be noted that the baseline system using an SVM-based approach provided by the shared task organizers attains an F-score of 68.98% for subtask A, thus surpassing all of the participants [Mohammad et al. 2016b]. Summaries of the participant papers of SemEval-2016 shared task on Twitter stance detection are given in Table 2.

Table 2. Participant Papers of SemEval-2016 Shared Task on Twitter Stance Detection [Mohammad et al. 2016b]

Authors	**Approach**	**Features**	**Subtask**
[Mohammad et al. 2016b]	SVM, majority class (baselines)	Word ngrams (1-3 gram) and character ngrams (2-5 gram)	A&B
[Zarrella and Marsh 2016]	LSTM	Learned features based on word and phrase embeddings from tweets	A
[Wei et al. 2016]	CNN	Learned features based on word embeddings from Google News database	A&B
[Tutek et al. 2016]	Ensemble learning based on SVM, random forest, gradient boosting, and logistic regression	Lexical features (word and character ngrams and word embeddings) and task-specific features (based on counts, misspelled words, and hashtags)	A
[Augenstein et al. 2016b]	Autoencoder for feature extraction and logistic regression for stance detection	Learned feature vector and "does target appear in tweet" feature	B
[Wojatzki and Zesch 2016a]	SVM	Word ngrams, syntactic, stance lexicon, concept, and target-transfer features	A
[Igarashi et al. 2016]	Logistic regression and CNN	Features based on target sentiment, ngrams, crawled tweets for logistic regression, word embeddings learned from Wikipedia for CNN	A

[Vijayaraghavan et al. 2016]	CNN	Character and word-level features	A
[Patra et al. 2016]	SVM	Features based on bag-of-words for each target, sentiment lexicons, and dependency relations	A
[Krejzl and Steinberger 2016]	Maximum entropy classifier	Ngrams, hashtags, POS tags, tweet length, sentiment and domain stance dictionaries	A&B
[Dias and Becker 2016]	SVM	Word ngrams (unigrams and bigrams)	B
[Zhang and Lan 2016]	Logistic regression	Linguistic, topic model, word vector, similarity, sentiment, and tweet-specific features	A&B
[Elfardy and Diab 2016]	SVM	Word ngrams (1-3 gram), topic models, sentiment analysis, word categories, and frame semantics	A
[Liu et al. 2016b]	Random forest, SVM, decision tree and ensemble classifiers	Unigrams and word vectors (word2vec [Mikolov et al. 2013] and GloVe [Pennington et al. 2014])	A
[Misra et al. 2016]	Multinominal naïve Bayes, SVM, decision tree	Unigrams, bigrams, POS tags, dependency relations, word counts, sentiment lexicons	A
[Bøhler et al. 2016]	Voting classifier (based on linear regression and multinominal naïve Bayes classifiers), SVM	Word bigrams, character trigrams, and GloVe word vectors	A

4.2.2 Shared Task of Stance Detection in Chinese Microblogs at NLPCC-ICCPOL-2016. A stance detection competition similar to the SemEval-2016 is conducted for Chinese microblog texts (from Sina Weibo application) as described in [Xu et al. 2016b]. In this competition, two subtasks are described similar to the settings of the corresponding SemEval-2016 competition: (1) subtask A where a supervised stance detection system is expected using the provided stance-annotated microblog dataset for training purposes, and (2) subtask B where an unsupervised system is expected as only a set of unlabeled microblog texts is provided. For subtask A, 4,000 microblogs are manually labeled for five targets, and 75% of them are used as the training dataset while the remaining 25% of them are used as the test dataset. The details of the dataset used in this competition are presented in Table 6 of Section 6.2. Summaries of the published papers presenting the approaches of the participants of this shared task are given in Table 3.

Table 3. Participant Papers of the Shared Task of Stance Detection in Chinese Microblogs at NLPCC-ICCPOL-2016 [Xu et al. 2016b]

Authors	**Approach**	**Features**	**Subtask**
[Sun et al. 2016]	SVM	Lexical (ngrams, post length, theme and context words), morphological (POS tags), semantic (polarity, sentiment/ stance words), and syntactic (dependency and syntax trees) features	A
[Yu et al. 2016]	LSTM	Word embeddings and word ngrams	A

[Liu et al. 2016a]	SVM, naïve Bayes, random forest, k-nearest neighbors (kNN), ensemble (voting) classifier	Ngrams with TF-IDF as the weighting scheme, sentiment features (polarity and the ratio of sentiment words)	A
[Xu et al. 2016a]	Linear SVM, SVM with RBF kernel, random forest, AdaBoost, and ensemble classifier	Bag-of-word features with TF and TF-IDF schemes, para2vec, features based on LDA, LSA, LE, LPI, sentiment, and subjectivity	A

16 teams participate in subtask A while five teams participate in subtask B. The system achieving the highest score of 71.06% in F-score is reported to use separate classifiers for each target and used classifiers based on SVM and random forest. The features employed by this top-scoring system include unigram, Term Frequency-Inverse Document Frequency (TF-IDF), synonym, and character and word vectors. Other features utilized by the other participants are word bigrams and sentiment lexicons. It is observed that multiple classifiers are used by the participants and using high-performance sentiment analysis systems may not guarantee improved stance detection performance. The performance of the participants is quite lower for subtask B when compared with that of subtask A, as expected. The highest performing system achieves an average F-score of 46.87% for subtask B [Xu et al. 2016b].

4.2.3 Shared Task of Stance Detection in Spanish and Catalan Tweets at IberEval-2017. A subsequent competition similar to SemEval-2016 and NLPCC-ICCPOL-2016 shared tasks on stance detection is conducted within the course of the IberEval-2017 conference which is a shared task on stance and gender detection from tweets in Spanish and Catalan [Taulé et al. 2017]. The dataset used in the stance detection competition is presented in Table 6 of Section 6.2. Summaries of the participant papers of this shared task are provided in Table 4.

Table 4. Participant Papers of the Shared Task of Stance and Gender Detection in Tweets on Catalan Independence at IberEval-2017 [Taulé et al. 2017]

Authors	**Approach**	**Features**	**Subtask**
[Taulé et al. 2017]	Majority class, LDR (baselines)	Term weights	Spanish & Catalan
[Lai et al. 2017]	SVM, logistic regression, decision tree, random forest, multinominal naïve Bayes, ensemble learner combining these classifiers, majority voting	Stylistic (word and character ngrams, POS tags, lemmas), structural (hashtags/mentions, hashtag frequencies, uppercase words, punctuation marks, numbers of words and characters), contextual (language, URL) features	Spanish & Catalan
[García and Flor 2017]	SVM and ANN	TF-IDF vectors of unigram and hashtag features	Spanish & Catalan
[Vinayakumar et al. 2017]	RNN, LSTM, GRU, and logistic regression	Word embeddings	Spanish & Catalan

[González et al. 2017]	SVM, LSTM, CNN, multilayer perceptron	Character and word ngrams, word embeddings vectors, character one-hot vectors, and a sentiment lexicon feature	Spanish
[Barbieri 2017]	FastText	Word embeddings considering subword information	Spanish & Catalan
[Swami et al. 2017]	SVM	Character (1-3) and word (1-5) ngrams, and stance indicative words	Spanish & Catalan
[Wojatzki and Zesch 2017]	SVM, LSTM, and a decision tree based hybrid system	Word (1-3) ngrams, character (2-4) ngrams, and word embeddings	Spanish & Catalan
[Ambrosini and Nicolo 2017]	LSTM, bidirectional LSTM, CNN	Word embeddings	Spanish & Catalan

Commonly employed approaches by the participants include SVM, neural networks, and deep learning methods such as Long Short-Term Memory (LSTM) which is a particular type of RNN, while the most common features are ngrams and word embeddings. The best performing system for stance detection on Spanish tweets is based on an SVM-based approach with a combination of different features while the best performer on Catalan tweets is based on logistic regression. Two worst performing systems are based on deep learning methods. Two baselines provided by the organizers are classifiers based on majority class and Low Dimensionality Representation (LDR) [Taulé et al. 2017].

5 APPROACHES TO STANCE DETECTION

Stance detection studies can be classified in several different ways. For instance, as previously mentioned, the studies conducted up to 2013 are classified into three groups in [Hasan and Ng 2013] based on the content type (all of which are posts published at online forums) used in these studies. Nevertheless, especially after competitions like the related SemEval-2016 shared task (see Section 4.2.1), the research attention is diverted to debates (and other topics) in online microblog posts, and mostly in tweets. Therefore, stance detection studies, so far, are mostly performed on online debates and microblog posts but it can be argued that the latter type now dominates the related literature.

In this section, we present related work on stance detection by categorizing them based on the approach that they employ, instead of the content type. Almost all of the studies are classification approaches, which can be divided into three categories: (1) feature-based machine learning approaches, (2) deep learning approaches, and (3) ensemble learning approaches. Related studies utilizing these approaches are described in the rest of this section, after statistical and insightful information about all of them as provided below.

Temporal distribution of published papers included in this survey paper is presented in Table 5. The total number of papers is 129. The content of Table 5 clearly indicates that research on stance detection boosts especially after 2015 and there is still an increasing trend in the number of studies performed.

A word cloud showing the frequencies of the employed classification algorithms used in the related studies is presented in Figure 4. The names of these algorithms are extracted from the content of the corresponding papers. It should be noted that if several algorithms are utilized in a paper, the frequencies of all of these algorithms are increased by one, and if a single classifier is tested with different configurations in a paper, the frequency is increased only by one for that particular classifier.

Table 5. Temporal Distribution of Published Papers on Stance Detection

Publication Year	Number of Papers
2006 – 2010	5
2011 – 2014	8
2015 – 2016	38
2017 – 2019	78

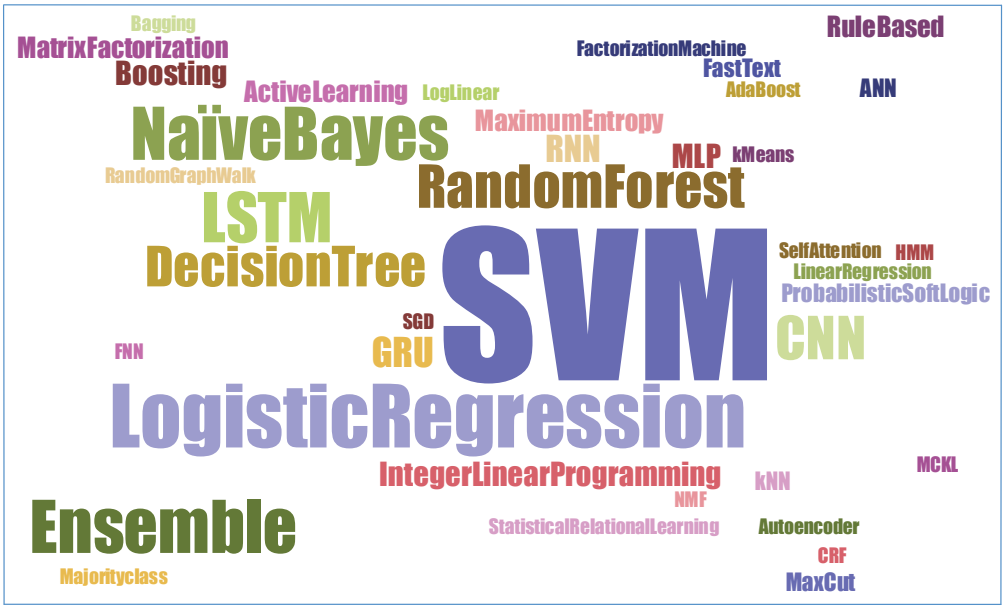


Fig. 4. A word cloud of the algorithms used for stance detection problem in the published papers included in this survey paper.

This word cloud demonstrates that traditional feature-based machine learning approaches like SVM, naïve Bayes, logistic regression, and decision tree algorithms are used more frequently than the other approaches in the stance detection literature, yet, deep learning methods (like LSTM and CNN) and ensemble methods including random forest algorithm are also utilized in a considerable number of studies.

During the evaluation of the presented approaches for stance detection, the datasets shared within the course of the related competitions are commonly used (such as [Mohammad et al. 2016b; Xu et al. 2016b]), in addition to the other available datasets (such as [Sobhani et al. 2017] for multi-target stance detection). Those studies conducted for rumour stance detection and fake news stance detection usually employ the corresponding shared datasets of these particular subproblems. In those studies on languages other than English, Chinese, Spanish, and Catalan (which are the languages of the shared datasets of the stance detection competitions), proprietary datasets are compiled and utilized. Section 6.2 of the current paper includes an overview of the stance detection datasets described and utilized in the related literature.

Before moving on to the reviews of the studies belonging to aforementioned three categories, we should note that there are few earlier studies on rule-based stance detection [Anand et al. 2011; Murakami and Raymond 2010; Walker et al. 2012a,b], as briefly covered previously in Section 4.1. All of these rule-based studies are reported to perform stance

detection in online debates. In [Murakami and Raymond 2010], a proprietary rule-based approach is employed where pattern dictionaries and the results of a sentiment analysis tool are used on text content in addition to the link structure in debates. In the remaining studies [Anand et al. 2011; Walker et al. 2012a,b], the rule-based JRip classifier is employed together with features based on ngrams, punctuations, dependencies, cue words, and post lengths, among others. Naïve Bayes is also tested in [Anand et al. 2011] and it is reported to outperform the rule-based JRip classifier. Due to the inherent limitations of the rule-based approaches for several NLP tasks including stance detection, learning approaches in the form of these basic three categories currently dominate stance detection studies.

In the following subsections, details of the related studies belonging to the relevant categories are provided. It should be noted that we do not repeat those studies that are already summarized in Table 2, 3, and 4 of Section 4.

5.1 Feature-based Machine Learning Approaches

It is a common practice for stance detection studies based on traditional feature-based machine learning approaches to employ and test more than one of approaches and compare them with each other. This pattern can well be observed in the studies participating in the related stance detection competitions, as demonstrated in Table 2, 3, and 4. Hence, while reviewing these approaches in this subsection, some studies will appear repeatedly under the discussions of different algorithms.

SVM is by far the most commonly employed feature-based machine learning approach for stance detection. SVMs are used in more than 40 studies on stance detection, either as the main best-scoring approach or as the baseline approach against which other approaches are compared. These studies include [Addawood et al. 2017; Bar-Haim et al. 2017; Dey et al. 2017; Gadek et al. 2017; Grčar et al. 2017; HaCohen-kerner et al. 2017; Hercig et al. 2017; Küçük 2017a,b; Küçük and Can 2018; Kucher et al. 2017; Lai et al. 2018; Mohammad et al. 2017; Rohit and Singh 2018; Sen et al. 2018; Siddiqua et al. 2018; Simaki et al. 2017a; Skeppstedt et al. 2016; Sobhani et al. 2015, 2016; Swami et al. 2018; Tsakalidis et al. 2018; Wojatzki and Zesch 2016b]. As reviewed in Section 4, SVMs are used both in earlier work as well as in stance detection competitions. For instance, in SemEval-2016 shared task [Mohammad et al. 2016b], baseline systems are based on SVMs and these baselines outperform all of the participating approaches. SVM-based participating systems are also reported to perform successfully in NLPCC-ICCPOL-2016 [Xu et al. 2016b] and IberEval-2017 [Taulé et al. 2017] shared tasks on stance detection (see Section 4.2 and Table 2, 3, and 4).

Logistic regression is the second most frequent classifier used for stance detection, appearing in more than 15 on-topic studies that we come across. In addition to those already mentioned in the previous section, some of the other studies using logistic regression for stance detection are [Ferreira and Vlachos 2016; HaCohen-kerner et al. 2017; Kucher et al. 2018; Lozhnikov et al. 2018; Purnomo et al. 2017; Sasaki et al. 2016; Simaki et al. 2017a; Skeppstedt et al. 2017; Tsakalidis et al. 2018; Zhang et al. 2017]. Similar to SVMs, logistic regression is known to perform favorably for stance detection task and is used either as the sole classifier or part of an ensemble classifier in related studies and competitions.

Considering the related literature that we cover in this paper, the probabilistic classifier, **naïve Bayes** is the third widely-employed algorithm of the traditional feature-based learning genre, appearing in more than 10 related studies. Some of these studies (excluding the ones already mentioned in Section 4 and Table 2, 3, and 4) are presented in [Addawood et al. 2017; HaCohen-kerner et al. 2017; Lai et al. 2016; Simaki et al. 2017a].

Next come **decision tree** classifiers, which appear in 9 studies on automatic stance detection such as [Addawood et al. 2017; HaCohen-kerner et al. 2017; Simaki et al. 2017a; Wojatzki and Zesch 2016b]. Random forest classifiers based

on decision trees are ensemble classifiers and they are used more frequently in stance detection studies compared to decision trees, as will be revisited in Section 5.3.

ANN is also employed in several related studies including [Sen et al. 2018; Tsakalidis et al. 2018]. Particularly, classifiers based on Multilayer Perceptron (**MLP**) are [Rajendran et al. 2018; Simaki et al. 2018; Zhang et al. 2018] successfully applied to the stance detection task.

Other traditional machine learning algorithms that are observed in stance detection literature are **ILP** [Ghosh et al. 2018; Konjengbam et al. 2018; Li et al. 2018], **kNN** [Shenoy et al. 2017], **log-linear model** [Ebrahimi et al. 2016a], **maximum entropy** [Hercig et al. 2017; Xu et al. 2017], **FastText** [Rohit and Singh 2018], Stochastic Gradient Descent (**SGD**) [Lozhnikov et al. 2018], **k-means clustering** [Simaki et al. 2017a], **matrix factorization** [Lin et al. 2017; Qiu et al. 2015; Sasaki et al. 2017], **factorization machines** [Sasaki et al. 2018], Multiple Convolution Kernel Learning (**MCKL**) [Tsakalidis et al. 2018], **statistical relational learning** [Ebrahimi et al. 2016b], and a **weakly-guided learning** scheme [Dong et al. 2017].

It should also be noted that some researchers employ **active learning** with the aforementioned frequently used classifiers for stance detection. For instance, in [Kucher et al. 2017; Skeppstedt et al. 2016] active learning with SVM is used for stance detection, and in a following study [Skeppstedt et al. 2017], active learning with a logistic regression classifier is used to detect cue words for stance/sentiment categories.

We conclude this subsection with the following list of common features utilized by the learning algorithms covered so far.

- Lexical features such as bag-of-words, word and character ngrams and skip-grams, hashtags, stance indicative words, theme and context words, synonyms, punctuation marks, and post length;
- Features based on interactions among posts and users (retweets, replies, agreement/disagreement links, quotes, geographic proximities, etc.) and temporal information regarding the posts;
- Features based on sentiment, subjectivity, and arguing/argumentation lexicons, emotion indicator words, and outputs of the related taggers;
- Word vector representations such as word2vec [Mikolov et al. 2013] and GloVe [Pennington et al. 2014] vectors (word embeddings), and paragraph vector representations such as para2vec [Le and Mikolov 2014];
- Topic modeling related features such as those based on Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and TF-IDF vectors of lexical features;
- Features based on POS tags, named entities, dependency relations, syntactic rules, and coreference resolution.

5.2 Deep Learning Approaches

Deep neural networks (such as RNNs with its modified versions, and CNNs) are employed in a considerable number of studies on stance detection. In several studies, it is a common practice to test a number of deep learning methods along with several traditional feature-based methods of the previous section and to compare the performance rates of these genres with each other. Therefore, some studies cited in this section are already cited in the previous section.

To begin with, **LSTM** [Hochreiter and Schmidhuber 1997], which is a type of RNN, is the most frequent deep learning approach used for stance detection, as revealed in more than 10 studies. These studies usually report that LSTMs perform favorably for this task. Apart from the ones already covered in Section 4.2, these studies include [Augenstein et al. 2016a; Dey et al. 2018; Du et al. 2017; Mavrin 2017; Rajendran et al. 2018; Sun et al. 2018a,b; Wei et al. 2018a]. Considering the same family of neural networks, about five studies report their related experiments with **RNN** including [Benton

and Dredze 2018; Mavrin 2017; Rajendran et al. 2018; Sobhani et al. 2017], and in six studies including [Benton and Dredze 2018; Hiray and Duppada 2017; Rajendran et al. 2018; Wei et al. 2018b; Zhou et al. 2017], Gated Recurrent Unit (**GRU**) [Chung et al. 2014] (another type of RNN) is employed as the main or baseline method, or as part of an ensemble method.

(**CNN**) is the second most frequent deep learning approach applied to stance detection, surpassing in number those studies based on RNNs and GRUs. Studies based on CNNs include [Hercig et al. 2017; Zhang et al. 2017; Zhou et al. 2017] in addition to the ones already covered in Section 4.2.

Some of the common features used by the related deep learning methods are word vector representations such as word2vec [Mikolov et al. 2013] and GloVe [Pennington et al. 2014] vectors (word embeddings) usually trained on large databases such as Google News database, phrase embeddings, word and character ngrams, and features based on sentiment lexicons.

As a concluding remark for this subsection; in many, and particularly recent, stance detection studies based on deep learning, an attention mechanism is introduced into the corresponding approach and it is reported to improve the stance detection performance [Dey et al. 2018; Du et al. 2017; Mavrin 2017; Sobhani et al. 2017; Sun et al. 2018b; Wei et al. 2018b; Xu et al. 2018; Zhou et al. 2017].

5.3 Ensemble Learning Approaches

Ensemble learning approaches for stance detection include those proposals in which more than one classifier are consolidated to arrive at a final stance output. They range from simpler combination schemes such as **majority voting** [Siddiqua et al. 2018] to more sophisticated approaches combining numerous different and successful classifiers.

To start with, random forest is an ensemble learning algorithm that combines several decision trees to cover the training dataset. **Random forest** algorithm is known to be one of the most frequent and effective ensemble learning algorithms for stance detection, as demonstrated in about 10 studies in the related literature [HaCohen-kerner et al. 2017; Shenoy et al. 2017; Swami et al. 2018; Tsakalidis et al. 2018], in addition to the participant systems of the stance detection competitions reviewed in Section 4.2.

Proprietary ensemble learners based on different number and type of learners are also frequently employed for stance detection, as observed in the ensemble-based participant systems of the stance detection competitions. Other studies that utilize ensemble learners for stance detection (or related subtasks such as debate detection) include [Zhang et al. 2017] where a combination of LSTM and CNN is used for the detection of debates, [Fraisier et al. 2018] where a semi-supervised ensemble algorithm is used, and [Rajendran et al. 2018] where combinations of LSTM and GRU are tested for stance detection although bidirectional LSTM outperforms these combinations. In [Zhou et al. 2017], a combination of bidirectional GRU and CNN with an attention mechanism is reported to outperform the SVM baseline (and the best performing approach) of SemEval-2016 shared task (Section 4.2.1) on the shared dataset. Similarly in [Wei et al. 2018b], an approach based on two bidirectional GRUs with a target-guided attention mechanism is employed which also outperforms the first-ranked SVM-based approach of SemEval-2016 shared task.

Other ensemble learners for stance detection include **boosting** and **bagging**, where related experiments are reported in [Lozhnikov et al. 2018; Simaki et al. 2017a].

Overall, the number of studies presenting ensemble learners for stance detection is considerably lower than those ones presenting traditional feature-based machine learning and deep learning. Yet, due to the significant potential of ensemble learners for diverse NLP problems, we believe that further comparative studies should be carried out in order

to firmly reveal whether ensembles of learners will boost stance detection performance compared to single learners, or not.

6 ANNOTATION GUIDELINES, DATASETS, AND EVALUATION METRICS

Stance detection is a considerably recent research topic and shared datasets with accompanied metrics and annotation guidelines are required in order to boost both the number and comparability of the related studies. In this section, we first review annotation guidelines for creating stance-annotated datasets, as described in the related literature. Next, we provide pointers to stance detection studies in the course of which related datasets are created. Finally, we describe common evaluation metrics used in stance detection studies.

6.1 Annotation Guidelines

Guidelines for stance annotation are usually provided in studies that describe stance detection competitions or those studies that take a linguistics-based point of view, as described below.

One of the most widely used datasets of stance detection is the dataset of English tweets created within the course of the SemEval-2016 shared task on stance detection (see Section 4.2.1 and Section 6.2). Guidelines provided to the annotators for the latest version of this dataset are described in [Mohammad et al. 2017]. This dataset is created through crowdsourcing (with the CrowdFlower tool) and the annotators are asked to answer two questions. The first question asks the stance class to be selected from one of the four classes: *Favor*, *Against*, *Neutral*, *Neither*. In order to clarify the scope of each class, possible cases that apply to the particular class are provided within the instructions. For instance, the *Favor* class can be selected if the tweet openly supports the target, or it supports an entity that is aligned with the target, or it opposes an entity from which it can be inferred that it supports the target, etc. In the second question, the annotators are asked to assess if the focus of the tweet is the stance target, or its focus is an entity other than the target, or whether it has a focus at all, or not. At the end of the annotation procedure, the number of tweets annotated with *Neutral* stance is found to be less than 1%, and therefore, the third and the fourth classes are combined into one stance class as *Neither* [Mohammad et al. 2017].

The dataset of Chinese microblogs, created for NLPCC-ICCPOL-2016 shared task on stance detection (see Section 4.2.2), contains annotations with one of the stance classes of *Favor*, *Against*, and *None* [Xu et al. 2016b]. The annotation is carried out by two students as annotators and if their stance annotations for a microblog do not coincide, then a third student is asked to classify it and the final stance class is determined by majority voting. The annotators are given a set of four instructions about the stance classes and how they should reason to arrive at the stance class when the stance target is not explicit and stance annotation is not straightforward [Xu et al. 2016b].

The dataset of Catalan and Spanish tweets compiled for the IberEval-2017 shared task on stance detection is annotated with one of the classes in {*Favor*, *Against*, *None*} by three annotators supervised by two researchers [Taulé et al. 2017]. The annotation is performed in three phases: (1) 500 tweets in each language are annotated, (2) inter-annotator agreement is calculated and possible inconsistencies are resolved, and (3) the annotators continue to annotate the whole dataset. During the evaluation of the annotation procedure, pairwise and average agreement percentages and Fleiss's Kappa coefficients are calculated [Taulé et al. 2017].

In [Simaki et al. 2017b], a corpus of blog posts annotated with cognitive/functional stance classes is described. The topic of the posts is the 2016 UK referendum regarding the Brexit event. 10 notional stance classes used to annotate this corpus are *Agreement/Disagreement*, *Certainty*, *Contrariety*, *Hypotheticality*, *Necessity*, *Prediction*, *Source of Knowledge*, *Tact/Rudeness*, *Uncertainty*, and *Volition*. Two annotators carry out the annotation procedure. A manual is provided to

them which includes information about the annotation process, the stance framework, and the annotation tool. They also participate in a related seminar given by a senior linguist and the annotation process start with a pilot round and is completed in two subsequent rounds where after the pilot round the annotators discuss their annotations with the linguist. The annotations are evaluated by calculating inter-annotator and intra-annotator agreement using the metrics of F-score and Cohen's Kappa coefficient [Simaki et al. 2017b].

6.2 Datasets

Although stance detection is a recent research topic, considerable effort is devoted to the creation of stance-annotated datasets, most of which are made publicly available. In the related literature, we come across stance detection datasets (of different text types such as tweets, posts in online forums, news articles, or news comments) for eleven languages: Arabic, Catalan, Chinese, Czech, English, English-Hindi, Italian, Japanese, Russian, Spanish, and Turkish. The details of the corresponding datasets, in terms of their domain, annotation classes, stance targets, sizes, and hyperlinks to access them (when applicable), are provided in Table 6. Earlier datasets mostly include online debate posts while more recent datasets include microblog posts like tweets. Almost all of the corresponding studies also report their stance detection experiments on these datasets, as previously reviewed in Section 5.

Table 6. Stance Detection Datasets

Authors	Domain	Annotation Classes	Target(s)	Size	URL
[Thomas et al. 2006]	Online political debates (English)	Yes, No	Proposed legislations	3,857 speech segments and 53 debates	[7]
[Somasundaran and Wiebe 2009]	Online debates on products (English)	Pro-Firefox, pro-IE, pro-iPhone, pro-Blackberry, pro-Opera, pro-Ps3, pro-Wii, pro-Windows, pro-Mac	Firefox vs. IE, iPhone vs. Blackberry, Opera vs. Firefox, Sony Ps3 vs. Nintendo Wii, Windows vs. Mac	304 debate posts	[8]
[Somasundaran and Wiebe 2010]	Online ideological debates (English)	For, Against (with topic level classes as Yes/No, Pro/Con etc.)	Several topics in healthcare, Existence of God, Gun rights, Gay rights, Abortion, and Creationism	7,134 debate posts	[9]
[Murakami and Raymond 2010]	Online debates (Japanese)	Support, Oppose	Selected five ideas	481 comments about five ideas	NA[10]

[7] http://www.cs.cornell.edu/home/llee/data/convote.html
[8] http://mpqa.cs.pitt.edu/corpora/product_debates/
[9] http://mpqa.cs.pitt.edu/corpora/political_debates/
[10] NA: Not applicable, i.e., not reported in the paper

[Levow et al. 2014]	Spontaneous speech (English)	No Stance, Weak Stance, Moderate Stance, Strong Stance, Unclear for stance; Positive, Negative, Neutral, Unclear for polarity	Decisions on item placement (inventory task) and whether to fund or cut expenses (budget task) in a superstore	~7.6 hours	NA
[Ferreira and Vlachos 2016]	Claims and news headlines (English)	For, Against, Observing	Claims extracted from rumour sites and Twitter	300 claims and 2,595 headlines	[11]
[Abbott et al. 2016]	Online debates (English)	Pro, Con	Various topics	482 posts	[12]
[Mohammad et al. 2016a]	Tweets (English)	Favor, Against, Neither	Atheism, Climate change is a real concern, Feminist movement, Hillary Clinton, Legalization of abortion, Donald Trump	4,870 tweets	[13]
[Mohammad et al. 2017]	Tweets (English)	Favor, Against, Neither for stance; Positive, Negative, and Neither for sentiment	Atheism, Climate change is a real concern, Feminist movement, Hillary Clinton, Legalization of abortion, Donald Trump	4,870 tweets	[14]
[Xu et al. 2016b]	Microblogs (Chinese)	Favor, Against, None	iPhone SE, Set off firecrackers in the Spring Festival, Russia's anti terrorist operations in Syria, Two child policy, Prohibition of motorcycles and restrictions on electric vehicles in Shenzhen, Genetically modified food, Nuclear test in DPRK	4,000 annotated and 2,400 unannotated tweets	NA
[Taulé et al. 2017]	Tweets (Catalan & Spanish)	Favor, Against, None	Independence of Catalonia	5,400 tweets in Spanish and 5,400 tweets in Catalan	[15]
[Sobhani et al. 2017]	Tweets (English)	Favor, Against, Neither	{Clinton-Sanders}, {Clinton-Trump}, {Cruz-Trump}	4,455 tweets	[16]

[11] https://github.com/willferreira/mscproject
[12] https://nlds.soe.ucsc.edu/iac2
[13] http://www.saifmohammad.com/WebPages/StanceDataset.htm
[14] http://www.saifmohammad.com/WebPages/StanceDataset.htm
[15] http://stel.ub.edu/Stance-IberEval2017/data.html
[16] http://www.site.uottawa.ca/~diana/resources/stance_data/

[Küçük 2017b]	Tweets (Turkish)	Favor, Against	Galatasaray, Fenerbahçe	700 tweets	[17]
[Küçük and Can 2018]	Tweets (Turkish)	Favor, Against	Galatasaray, Fenerbahçe	1,065 tweets	[18]
[Addawood et al. 2017]	Tweets (English)	Favor, Against, Neutral	Individual privacy, Natural security, Other, Irrelevant	3,000 tweets	NA
[Darwish et al. 2017]	Tweets (Arabic)	Favor (Positive), Against (Negative)	Transfer of two islands from Egypt to Saudi Arabia	33,024 tweets	NA
[Hercig et al. 2017]	News comments (Czech)	In Favor, Against, Neither	Miloš Zeman, Smoking ban in restaurants	5,423 news comments	[19]
[Derczynski et al. 2017]	Tweets (English)	Support, Deny, Query, Comment	Rumorous tweets	5,568 tweets (4,519 + 1,049)	[20]
[FNC 2017]	News headlines and body texts (English)	Agrees, Disagrees, Discusses, Unrelated	News headlines	49,972 annotated and 25,413 unannotated headline-body pairs	[21]
[Baly et al. 2018]	Web sites (Arabic)	Agree, Disagree, Discuss, Unrelated	Claims extracted from Web sites	402 claims and 3,042 annotated documents	[22], [23]
[Swami et al. 2018]	Tweets (English-Hindi)	Favor, Against, None	Demonetisation in India in 2016	3,545 tweets	[24]
[Lai et al. 2018]	Tweets (Italian)	Favor, Against, None	2016 referendum on reform of the Italian Constitution	993 triplets (2,889 tweets)	[25]
[Rohit and Singh 2018]	Online Indian debates (English)	Favor, Against for stance; Appreciate, Blame, Call for Action, Issue for purpose	The bill/issue of the speech under consideration	1,201 speeches	[26]

[17] https://github.com/dkucuk/Stance-Detection-Turkish-V1
[18] https://github.com/dkucuk/Stance-Detection-Turkish-V3
[19] http://nlp.kiv.zcu.cz/research/sentiment#stance
[20] https://s3-eu-west-1.amazonaws.com/downloads.gate.ac.uk/pheme/semeval2017-task8-dataset.tar.bz2
[21] https://github.com/FakeNewsChallenge/fnc-1
[22] http://groups.csail.mit.edu/sls/downloads/factchecking/
[23] http://alt.qcri.org/resources/
[24] https://github.com/sahilswami96/StanceDetection_CodeMixed
[25] https://github.com/mirkolai/Stance-Evolution-and-Twitter-Interactions
[26] The dataset is stored as a publicly available online database which can be accessed using the following command on Linux systems: mongo ds235388.mlab.com:35388/synopsis -u public -p public

[Lozhnikov et al. 2018]	Tweets and news (Russian)	Support, Deny, Query, Comment	Claims extracted from news and tweets	700 tweets and 200 news articles	[27]

6.3 Evaluation Metrics

The metrics of precision, recall, and F-score (or F-measure) are commonly used in information retrieval and information extraction research. According to the definitions based on the ones in [van Rijsbergen 1979], F-score (denoted as F in following formulae) is a combined metric calculated using precision (P) and recall (R) with the flexibility to give weights to these two metrics. Hence its generic form is provided as the first formula below. Most of the time, β is taken as 1, by giving equal weights to precision and recall, and F-score becomes the harmonic mean of them, as shown next to the generic formula below.

$$F = \frac{(\beta^2 + 1) * P * R}{\beta^2 * P + R} \quad (\beta \geq 0) \qquad F = \frac{2 * P * R}{P + R} \quad (\beta = 1)$$

This evaluation metric of F-score is also commonly employed in three-way stance detection (into one of the three classes: *Favor*, *Against*, and *Neither*). The most widely-used version of F-score is calculated as the macro-average of the F-scores for *Favor* and *Against* classifications as follows [Mohammad et al. 2016b; Taulé et al. 2017; Xu et al. 2016b]:

$$F = \frac{F_{Favor} + F_{Against}}{2} \qquad F_{Favor} = \frac{2 * P_{Favor} * R_{Favor}}{P_{Favor} + R_{Favor}} \qquad F_{Against} = \frac{2 * P_{Against} * R_{Against}}{P_{Against} + R_{Against}}$$

$$P_{Favor} = \frac{Correct_{Favor}}{Correct_{Favor} + Spurious_{Favor}} \qquad P_{Against} = \frac{Correct_{Against}}{Correct_{Against} + Spurious_{Against}}$$

$$R_{Favor} = \frac{Correct_{Favor}}{Correct_{Favor} + Missing_{Favor}} \qquad R_{Against} = \frac{Correct_{Against}}{Correct_{Against} + Missing_{Against}}$$

It is pointed out in the related literature that F-score calculated this way (as the macro-average of the calculations for two classes only) does not disregard the *Neither* class, since incorrectly classifying content as *Neither* (instead of *Favor* or *Against*) affects the calculated score.

In some other studies such as [Hercig et al. 2017], the macro-average of the scores of all three classes are also used for evaluation as follows:

$$F = \frac{F_{Favor} + F_{Against} + F_{Neither}}{3}$$

Accuracy is another metric used for stance detection [Hercig et al. 2017] and is calculated as follows:

$$Accuracy = \frac{Correct\ classifications}{All\ classifications}$$

[27] https://github.com/npenzin/rustance

In [Murakami and Raymond 2010], accuracy is defined as the average of the accuracies for *Support* and *Oppose* classifications, in order to reduce the potential bias due to the possibility that the dataset is not well-balanced. In the corresponding formula below, A and S are the classifications in the answer key and system output respectively, and *Sup* and *Opp* denote *Support* and *Oppose*, respectively.

$$Accuracy = \frac{1}{2} * \left(\frac{|A_{Sup} \cap S_{Sup}|}{|A_{Sup}|} + \frac{|A_{Opp} \cap S_{Opp}|}{|A_{Opp}|} \right)$$

7 SOFTWARE AND TOOLS

Software and tools related to the topic of stance detection can be categorized into two groups: (1) applications which include a stance detection module, and (2) generic machine learning platforms/tools/libraries commonly used to train stance detection models.

The number of studies in the first group is rather low, we come across only four papers describing such systems. These studies are overviewed below.

- A stance community detection application, called SCIFNET, is described in [Chen and Chen 2016] which clusters topic persons in documents into communities.
- In [Ruder et al. 2018], a tool is described which collects online news articles, performs stance detection of these articles towards a precompiled set of topics (popular, political, and controversial) using the bidirectional conditional encoding approach in [Augenstein et al. 2016a], and visualizes the results with links to original news sources.
- In [Kucher et al. 2017], a visual analysis and active learning system for stance detection, called ALVA, is presented. The stance detection approach utilized is SVM with active learning facility.
- By the same authors, a system for visualizing and annotating stance and sentiment information, called StanceVis Prime, is described in [Kucher et al. 2018].

A list of generic machine learning tools used to develop the stance detection approaches reviewed in Section 5, which constitute the aforementioned second group of related software and tools, is provided below with references to the corresponding stance detection studies as well.

- SVM implementation in **Weka toolkit** [Hall et al. 2009] is employed for stance detection in [Elfardy and Diab 2016; Küçük 2017a,b; Küçük and Can 2018; Somasundaran and Wiebe 2010; Wojatzki and Zesch 2016a]. SVM, naïve Bayes, and J48 (decision tree) classifiers in Weka [Hall et al. 2009] are also used in [Addawood et al. 2017; Misra et al. 2016]. Random forest and J48 classifiers in [Misra and Walker 2013], and naïve Bayes, JRip classifiers in [Anand et al. 2011; Walker et al. 2012b] also use the corresponding implementations in Weka [Hall et al. 2009]. Weka is also used for all of the classifiers employed in the experiments reported in [Simaki et al. 2017a].
- [Rajadesingan and Liu 2014], [Augenstein et al. 2016b], [Tutek et al. 2016], [Liu et al. 2016b], [Bøhler et al. 2016], [Skeppstedt et al. 2016], [Skeppstedt et al. 2017], [Swami et al. 2017], [Ferreira and Vlachos 2016], [Lai et al. 2016], [Shenoy et al. 2017], [Mohammad et al. 2017], [Simaki et al. 2018] use the **scikit-learn package** [Pedregosa et al. 2011] for the feature-based machine learning approaches that they employ.
- **Keras library** [Chollet 2015] is used in [Zarrella and Marsh 2016] for RNN implementation, and also in [Lozhnikov et al. 2018].
- **Theano library** [Theano Development Team 2016] is used in [Wei et al. 2016] for CNN implementation.

- **Gensim library** [Rehurek and Sojka 2010] is used in [Lozhnikov et al. 2018] for vector space and topic modelling representation.
- **SVM^{light}** implementation available at `http://svmlight.joachims.org/` is used in [Thomas et al. 2006].
- The source code of the **FastText** algorithm proposed in [Joulin et al. 2016] for text classification and utilized in [Barbieri 2017; Rohit and Singh 2018] for stance detection is available at `https://github.com/facebookresearch/fastText`.
- The implementations of maximum entropy and SVM classifiers in the **Brainy library** [Konkol 2014] are used in [Hercig et al. 2017] for stance detection in news comments.

Finally, some of the researchers of stance detection publicly share their related resources and source codes. Interested readers are referred to [Wei et al. 2016], [Augenstein et al. 2016b], [Liu et al. 2016a], [Xu et al. 2016a], and [Riedel et al. 2017] which are among such studies with public links to the corresponding stance detection resources and source codes. Such studies are significant as they enable the replication of the experiments described by other researchers and hence facilitate comparisons with new proposals. Therefore, we believe that commonly sharing source codes of the approaches described in upcoming stance detection studies will lead to improved stance detection performance by the studies following them.

8 APPLICATION AREAS

Stance detection is known to have a diverse set of application areas. One of the most common of them is **opinion surveys/polling**. Stance detection studies are carried out mostly on online textual content where the topics include political/ideological/social debates, product reviews, and elections/referendums. Hence, by means of automatic stance detection, whether a community is in favor of or against a topic of interest can be estimated, replacing (or, complementing) the traditional practices of performing surveys/polls. In a similar fashion, stance detection can also be utilized to facilitate **trend and market analysis/forecast** by using the evolution of community stance in time. Thirdly, **recommendation systems** can benefit from the stance detection patterns of individuals to provide them with more personalized recommendations.

On the other hand, stance detection is also applied to facilitate **public health surveillance**, as exemplified in [Zhang et al. 2017] where the authors perform the following tasks on online health forums: (1) identification of controversial discussions regarding complementary and alternative medicine (CAM), (2) stance detection on the posts in these discussions, and (3) manual identification of the CAM therapies likely to trigger debates. In another study [Purnomo et al. 2017], the logistic regression based classifier in [Ferreira and Vlachos 2016] is employed for automatic detection of hoax medical news articles (similar to fake news detection) using stance detection. Similarly in [Sen et al. 2018], the authors focus on health information retrieval with stance-based categorization of the search results. Hence, another plausible application area for stance detection is **information retrieval**. Information retrieval systems with a focus on stance-bearing content [Pariser 2011] can be developed in order to provide targeted content with convenient visualization facilities to interested users. For instance, the sentiment retrieval system described in [Miao et al. 2009] has proprietary ranking and visualization features for product reviews, similar systems can be developed for online content from the perspective of stance. Another application area, also related to information retrieval, is **stance summarization**. In a related study [Jang and Allan 2018], the authors present an unsupervised approach to produce stance-aware summaries of controversies in Twitter, by identifying a (ranked) tweet set that best represents the controversy under discussion.

In addition to the six areas mentioned above, two other significant application areas of stance detection are **rumour classification** and **fake news detection**. The systems that address these topics include stance detection modules tailored to their particular needs and, as presented in the last two definitions given in Section 1, the stance-related problems in these topics can be considered as distinct subproblems of stance detection. In order to review the related literature adequately, these two application areas are discussed in the following separate subsections.

8.1 Rumour Classification

Especially with the widespread use of social media, rumours start to circulate quickly, and this phenomenon calls for automatic ways to identify and resolve these rumours. A survey of rumour classification (or, rumour detection and resolution) studies on social media is presented in [Zubiaga et al. 2018a], where a rumour is defined as a piece of information that has not yet been verified. A rumour classification system is reported to have four basic components: two components for rumour identification and tracking, respectively, and two components for the classification of rumour stance and its veracity, respectively [Zubiaga et al. 2018a]. Hence, it is the third component of the system where stance detection comes into play. As previously defined in Section 1, rumour stance classification aims to determine the orientation of a given post with respect to a given rumour, usually as a class label from this set: {*Supporting*, *Denying*, *Querying*, *Commenting*}. The output of the rumour stance classification module is used by the last (veracity classification) module of the overall system, which produces a veracity class label for the given rumour, usually from this set: {*True*, *False*, *Unverified*} [Zubiaga et al. 2018a].

Rumour stance detection is a recent and popular research topic, and therefore, rumour classification is an important application area of stance detection. Similar to the approaches for stance detection, those for rumour stance detection are usually supervised machine learning approaches with different feature sets [Lukasik et al. 2019; Pamungkas et al. 2019; Zubiaga et al. 2018a, 2016, 2018b] in addition to semi-supervised approaches [Giasemidis et al. 2018]. Approaches based on deep learning methods [Zubiaga et al. 2018b] and those additionally utilizing attention mechanisms [Veyseh et al. 2017] are employed for rumour stance detection as well. There are also annotated and public datasets regarding rumour stance (such as [Ferreira and Vlachos 2016]) and competitions with associated datasets (such as [Derczynski et al. 2017]), as given in Table 6 of Section 6.2.

8.2 Fake News Detection

Similar to the case of rumours, fake news constitutes another source of misinformation online. Accordingly, fake news detection has emerged as an important research problem recently, which aims at determining fake news published in online information channels. Fake news is defined in [Lazer et al. 2018] as fabricated information that seems like genuine news content, but the creation of which lacks the required norms and processes to ensure its accuracy and credibility.

An important milestone for fake news detection, and the use of stance detection within the solution of this problem, is the related competition known as the *Fake News Challenge* (FNC) [FNC 2017]. In FNC, stance detection is considered as a useful first stage (also referred to as FNC-1) to determine whether a given story is real or fake. Fake news stance detection within the context of FNC is previously defined in Section 1 (as Definition 1.5) which is based on the corresponding definition in [Ferreira and Vlachos 2016]. According to this definition, in fake news stance detection, given a headline and a body text, the stance of the body text with respect to the headline is expected from this set: {*Agrees*, *Disagrees*, *Discusses*, *Unrelated*}. Within the course of FNC, annotated training and unannotated test datasets are made publicly available (see Table 6 of Section 6.2).

Approaches to fake news stance detection also come in the form of different learning systems with various feature sets, similar to the approaches for generic stance detection. The baseline system in FNC is a gradient boosting classifier which is reported to attain a weighted accuracy of 79.53%. There is a significant body of work on fake news stance detection including both the participants of FNC and the studies performed after the competition, including but not limited to [Bhatt et al. 2018; Ghanem et al. 2018; Hanselowski et al. 2018; Masood and Aker 2018; Riedel et al. 2017; Shang et al. 2018; Shu et al. 2017; Thorne et al. 2017; Yang et al. 2019].

As a final note, **automatic fact checking** is a closely-related problem and arguably a generalized form of fake news detection. Fact checking is defined in [Vlachos and Riedel 2014] as the procedure of determining the truth value of a claim made in a given context. Accordingly, stance detection within the context of automatic fact checking is treated similarly to fake news stance detection in studies such as [Mohtarami et al. 2018] where different forms of deep neural networks are tested on the FNC dataset during the evaluation procedure. The authors of [Mohtarami et al. 2018] are also involved in the compilation of a stance detection dataset for fact checking in [Baly et al. 2018], similar to the FNC dataset, as previously summarized in Table 6.

9 OUTSTANDING ISSUES WITH FUTURE RESEARCH POSSIBILITIES

There are several outstanding issues regarding stance detection, which correspond to a number of future research possibilities based on the existing literature. These future research topics include: (1) cross-lingual and multilingual stance detection, (2) stance detection in other media content and robots, (3) stance detection for decision making, (4) stance detection in data streams, (5) stance detection and deep NLP, (6) issues regarding datasets and evaluation, and (7) context-sensitive stance detection. These topics are discussed in the following subsections below.

9.1 Cross-lingual and Multilingual Stance Detection

As is the case for several topics in NLP, stance detection necessitates annotated datasets so that related research experiments can be conducted for different languages. Stance detection datasets are produced for eleven languages so far, most of which are made publicly available (see Section 6.2). Yet, most of these datasets are in English and they are far from sufficient for extensive multilingual stance detection experiments.

Within the course of future studies, to perform stance detection studies in languages that lack annotated datasets, **cross-lingual stance detection** can be employed as follows: annotated dataset in a given language (e.g., English) can be automatically translated into the target languages, and using the translations (together with stance labels) as training datasets, related classifiers/models can be built which can then be tested on smaller test datasets in these languages. Such experiments for the task of cross-lingual topic tracking are described in [Allan et al. 2003]. Hence, we believe that cross-lingual stance detection will be a fruitful line of future work. Similarly, **multilingual stance detection** appears as a plausible future research topic where stance detection is performed on content available in several different languages. Related studies for sentiment analysis can be found in [Boiy and Moens 2009; Can et al. 2018]. We note that differences in the linguistic phenomena governing the languages under consideration should be carefully analyzed when performing cross-lingual and multilingual stance detection. Insights from related studies on sentiment analysis such as [Boyd-Graber and Resnik 2010] can be used during cross-lingual and multilingual stance detection.

9.2 Stance Detection in Other Media Content and Robots

Almost all of the related research on stance detection that we come across is carried out on textual content, the only exception being the study by [Levow et al. 2014] which focuses on stance detection on speech. Yet, stance detection

on other content, such as images and videos, is a promising line of future work. Research on computer vision and particularly on semantic content extraction from images and videos is known to make an accelerated progress recently, together with the widespread use of deep learning methods. Automatic and joint stance detection from these different modalities can contribute to the creation and enrichment of automatic descriptions and summaries of images/videos which are performed traditionally through the extraction of objects and events from the visual and audio content.

Stance detection through visual or speech analysis (in addition to the textual analysis) will ultimately contribute to the goal of designing and building emotional robots. Chatbots which can perform argumentative conversations by taking different stances than humans, such as Debbie [Rakshit et al. 2019], constitute an important step towards this aim. Such chatbots of the future can be built to perform stance analysis in text, speech, image, and video when they engage in human-like conversions with their users. In [Breazeal and Brooks 2005], it is pointed out that emotions play an important role in several processes of humans such as decision making, planning, and learning. Therefore, emotion-inspired capabilities will be crucial during the design of future autonomous robots, particularly for improved and effective interaction with humans [Breazeal and Brooks 2005]. Accordingly, stance detection and emotion recognition (described in Section 2), can be important features of prospective emotional robots.

9.3 Stance Detection for Decision Making

A topic closely related to the open issues covered in the previous subsection is the use of stance detection for decision support. It is emphasized in the aforementioned study on emotional robots [Breazeal and Brooks 2005] that decision making is one of the significant capabilities of intelligent creatures which make use of both cognition and emotion. Hence, the results of automatic stance detection can be used to aid in the decision making processes of both humans and autonomous robots alike. For instance, humans can benefit from the outcomes of stance-based information retrieval systems (such as [Sen et al. 2018]) which present and visualize the frequency and temporal evolution of different stances. Similarly, humanoid robots can utilize the community stance when making decisions regarding its operations in different settings. Recent online ensemble methods such as [Büyükçakir et al. 2018] can be used to aggregate the outputs of different classifiers used for community stance detection.

9.4 Stance Detection in Data Streams

Data streams are usually defined as large volumes of data that are retrieved continuously, which require almost real-time processing capabilities to analyze the incoming data adequately and store the analysis results if necessary [Muthukrishnan 2005]. Data stream processing capabilities are also useful for ever-increasing online textual content, such as microblogs or online debate posts. For instance, in [Bifet and Frank 2010], challenges of sentiment analysis on Twitter streaming data are discussed. In a similar fashion, it will be a significant direction of future research to perform stance detection in streaming online content, so that almost real-time stance results can be obtained which can be used in different application settings covered in Section 8.

9.5 Stance Detection and Deep NLP

Common features used in stance detection studies so far are based on ngrams, word embedding vectors, sentiment lexicons, hashtags, and term frequencies, among others. While few studies also utilize features based on POS tags, syntax trees, and dependencies [Shenoy et al. 2017; Sun et al. 2016], there is a need for studies that will assess the possible contribution of using deeper language processing to stance detection. It is expected that a computational cost will be introduced due to the time and space complexities of these language processing schemes like full parsing which

can hinder their employment in real-time application settings, but related future work can help reveal whether the contribution that they provide is worth the cost that they introduce into the stance detection pipelines.

A recent trend in language processing is the employment of language representation models for pretraining on large unannotated corpora like Wikipedia before they can be fine-tuned on domain-specific (and most of the time, limited) annotated corpora [Devlin et al. 2018]. Google's Bidirectional Encoder Representations from Transformers (BERT) is one such model which is contextual and bidirectional as it uses both previous and following contexts of words [Devlin et al. 2018]. BERT is reported to lead to considerable performance improvements for NLP tasks such as sentiment analysis [Devlin et al. 2018]. Hence, another plausible direction of future work is the use and assessment of such models for stance detection considering both theoretical and practical aspects.

9.6 Issues Regarding Datasets and Evaluation

Another future work direction is in-depth analysis of the annotated datasets, annotation guidelines, and evaluation metrics and results presented so far, in terms of reliability and effectiveness. A study which targets at the analysis of some of these issues for information retrieval is presented in [Zobel 1998]. An issue related to tweet datasets is that such datasets are shared with tweet identifiers only, instead of the actual contents (due to the constraints imposed by Twitter), and at the time of a new study to be carried out on such shared datasets, some referenced tweets may not be available due to subsequent deletions by the posters. The authors of the current paper are contacted by different researchers due to this phenomena, as it is observed in the stance-annotated tweet dataset that they publicly share [Küçük and Can 2018].

There is also a need to compile larger annotated datasets, preferably of different text genres and in different languages, in order to increase the performance and applicability of state-of-the-art stance detection approaches.

An important related issue is the need for statistical tests to validate the significance of the attained results. Hence, it is expected from prospective stance detection studies to apply convenient statistical tests to validate their results and make it a common practice to do so.

9.7 Context-sensitive Stance Detection

Modeling context in the forms of spatial and temporal locality is known to be crucial in diverse application domains, including memory management systems, search engines, and context-aware Web applications [Denning 2005].

In his work [Du Bois 2007] on a linguistic framework for stance interpretation, Du Bois claims that a context-free interpretation of stance, considering only an isolated single sentence, will be incomplete. Similarly, from an application-oriented point of view, stance detection is expected to benefit from a context-sensitive approach, as demonstrated by related studies which make use of conversational or dialogic interactions (such as retweets, replies), and user modeling during stance detection [Lai et al. 2018; Li et al. 2018; Sasaki et al. 2018]. Context-sensitive approaches are also proposed for related problems like sentiment analysis [Ren et al. 2016]. Therefore, a significant future work direction could be the exploitation of context for improved stance detection, as performed in recent studies such as [Lukasik et al. 2019; Veyseh et al. 2017]. Context-sensitive stance detection on social media can utilize other online content produced by the user through time, can process and detect stance in the content produced by other users with whom the user interacts with. It is expected that, when more contextual information is introduced into the stance detection procedure, the performance rates of the procedure will be improved.

10 CONCLUSIONS

Stance detection is usually defined as the automatic determination of the position of a post owner (as in favor of or against) towards a specific target, based on the content of the post. Along with a number of related problems such as sentiment analysis, controversy detection, and argument mining, it is a crucial process to elicit useful information from the underlying content, most of the time, regarding controversial issues or elections/referendums. In this paper, we present a comprehensive survey of automatic stance detection studies. In addition to providing related definitions and describing its related topics, the current paper presents the related studies as categorized by the approach employed. A generic system architecture for stance detection, related annotation guidelines, datasets, metrics, application areas, and outstanding issues are also included in the paper. We believe that the paper will be beneficial to NLP researchers who need to learn about the state-of-the-art regarding stance detection. It will also be useful for NLP practitioners who would like to build stance-oriented automatic information elicitation systems for the vast amount of textual data that is publicly available online. In other words, this comprehensive survey will help researchers and practitioners untangle the background and skeleton of stance detection, together with the presented insights to facilitate future research.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their insightful comments.

REFERENCES

2017. Fake news challenge stage 1 (FNC-1): stance detection. Retrieved 9 May 2018 from http://www.fakenewschallenge.org/

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn A Walker. 2016. Internet Argument Corpus 2.0: an SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Language Resources and Evaluation Conference*.

Aseel Addawood, Jodi Schneider, and Masooda Bashir. 2017. Stance classification of Twitter debates: the encryption debate as a use case. In *Proceedings of the 8th International Conference on Social Media & Society*. 2.

Mahmoud Al-Ayyoub, Abdullateef Rabab'ah, Yaser Jararweh, Mohammed N Al-Kabi, and Brij B Gupta. 2018. Studying the controversy in online crowds' interactions. *Applied Soft Computing* 66 (2018), 557–563.

James Allan, Victor Lavrenko, and Margaret E Connell. 2003. A month to topic detection and tracking in Hindi. *ACM Transactions on Asian Language Information Processing* 2, 2 (2003), 85–100.

Luca Ambrosini and Giancarlo Nicolo. 2017. Neural models for StanceCat shared task at IberEval 2017. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*.

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*. 1–9.

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

Isabelle Augenstein, Andreas Vlachos, and Kalina Bontcheva. 2016b. USFD at SemEval-2016 task 6: any-target stance detection on Twitter with autoencoders. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 389–393.

Ramy Baly, Mitra Mohtarami, James Glass, Lluis Marquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of NAACL-HLT*.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 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. 251–261.

Francesco Barbieri. 2017. Shared task on stance and gender detection in tweets on Catalan independence-LASTUS system description. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*.

Adrian Benton and Mark Dredze. 2018. Using author embeddings to improve tweet stance classification. In *Proceedings of the EMNLP Workshop on Noisy User-generated Text (W-NUT)*. 184–194.

Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. 2018. Combining neural, statistical and external features for fake news stance identification. In *Proceedings of the Web Conference Companion*. 1353–1357.

Albert Bifet and Eibe Frank. 2010. Sentiment knowledge discovery in twitter streaming data. In *International Conference on Discovery Science*. 1–15.

Henrik Bøhler, Petter Asla, Erwin Marsi, and Rune Sætre. 2016. IDI@NTNU at SemEval-2016 task 6: detecting stance in tweets using shallow features and GloVe vectors for word representation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 445–450.

Erik Boiy and Marie-Francine Moens. 2009. A machine learning approach to sentiment analysis in multilingual Web texts. *Information Retrieval* 12, 5 (2009), 526–558.

Hamed Bonab and Fazli Can. 2018. GOOWE: geometrically optimum and online-weighted ensemble classifier for evolving data streams. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 12, 2 (2018), 25.

Jordan Boyd-Graber and Philip Resnik. 2010. Holistic sentiment analysis across languages: multilingual supervised latent Dirichlet allocation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 45–55.

Cynthia Breazeal and Rodney Brooks. 2005. Robot emotion: A functional perspective. *Who needs emotions* (2005), 271–310.

Alican Büyükçakir, Hamed Bonab, and Fazli Can. 2018. A novel online stacked ensemble for multi-label stream classification. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 1063–1072.

Ethem F Can, Aysu Ezen-Can, and Fazli Can. 2018. Multilingual sentiment analysis: an RNN-based framework for limited data. *arXiv preprint arXiv:1806.04511* (2018).

Zhong-Yong Chen and Chien Chin Chen. 2016. SCIFNET: stance community identification of topic persons using friendship network analysis. *Knowledge-Based Systems* 110 (2016), 30–48.

François Chollet. 2015. Keras. https://keras.io/

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Improved stance prediction in a user similarity feature space. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. 145–148.

Peter J Denning. 2005. The locality principle. *Commun. ACM* 48, 7 (2005), 19–24.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2017. Twitter stance detection–a subjectivity and sentiment polarity inspired two-phase approach. In *Proceedings of the ICDM Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction (SENTIRE)*.

Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2018. Topical stance detection for Twitter: a two-phase LSTM model using attention. *arXiv preprint arXiv:1801.03032* (2018).

Marcelo Dias and Karin Becker. 2016. INF-UFGRS-OPINION-MINING at SemEval-2016 task 6: automatic generation of a training corpus for unsupervised identification of stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 378–383.

Edsger W Dijkstra. 1997. The tide, not the waves. In *Beyond calculation*. Springer, 59–64.

Rui Dong, Yizhou Sun, Lu Wang, Yupeng Gu, and Yuan Zhong. 2017. Weakly-guided user stance prediction via joint modeling of content and social interaction. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 1249–1258.

Shiri Dori-Hacohen. 2015. Controversy detection and stance analysis. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1057.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*.

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

Javid Ebrahimi, Dejing Dou, and Daniel Lowd. 2016a. A joint sentiment-target-stance model for stance classification in tweets. In *Proceedings of the International Conference on Computational Linguistics*. 2656–2665.

Javid Ebrahimi, Dejing Dou, and Daniel Lowd. 2016b. Weakly supervised tweet stance classification by relational bootstrapping. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1012–1017.

Heba Elfardy and Mona Diab. 2016. CU-GWU perspective at SemEval-2016 task 6: ideological stance detection in informal text. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 434–439.

Adam Faulkner. 2014. Automated classification of stance in student essays: an approach using stance target information and the Wikipedia link-based measure. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference*. 174–179.

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

Ophélie Fraisier, Guillaume Cabanac, Yoann Pitarch, Romaric Besançon, and Mohand Boughanem. 2018. Stance classification through proximity-based community detection. In *Proceedings of the ACM Conference on Hypertext and Social Media (HT)*.

Guillaume Gadek, Josefin Betsholtz, Alexandre Pauchet, Stéphan Brunessaux, Nicolas Malandain, and Laurent Vercouter. 2017. Extracting contextonyms from Twitter for stance detection. In *Proceedings of the 9th International Conference on Agents and Artificial Intelligence (ICAART)*. 132–141.

Diego Aineto García and Antonio Manuel Larriba Flor. 2017. Stance detection at IberEval 2017: a biased representation for a biased problem. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*.

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

Subrata Ghosh, Konjengbam Anand, Sailaja Rajanala, A Bharath Reddy, and Manish Singh. 2018. Unsupervised stance classification in online debates. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*. 30–36.

Georgios Giasemidis, Nikolaos Kaplis, Ioannis Agrafiotis, and Jason Nurse. 2018. A semi-supervised approach to message stance classification. *IEEE Transactions on Knowledge and Data Engineering* (2018).

José-Ángel González, Ferran Pla, and Lluıs-F Hurtado. 2017. ELiRF-UPV at IberEval 2017: stance and gender detection in tweets. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*.

Miha Grčar, Darko Cherepnalkoski, Igor Mozetič, and Petra Kralj Novak. 2017. Stance and influence of Twitter users regarding the Brexit referendum. *Computational Social Networks* 4, 1 (2017), 6.

Yaakov HaCohen-kerner, Ziv Ido, and Ronen Ya'akobov. 2017. Stance classification of tweets using skip char ngrams. In *Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 266–278.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter* 11, 1 (2009), 10–18.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance detection task. In *Proceedings of the International Conference on Computational Linguistics*. 1859–1874.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: data, models, features, and constraints. In *Proceedings of the International Joint Conference on Natural Language Processing*. 1348–1356.

Tomáš Hercig, Peter Krejzl, Barbora Hourová, Josef Steinberger, and Ladislav Lenc. 2017. Detecting stance in Czech news commentaries. In *Proceedings of the Conference on Theory and Practice of Information Technologies (ITAT)*.

Sushant Hiray and Venkatesh Duppada. 2017. Agree to disagree: improving disagreement detection with dual GRUs. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

Yuki Igarashi, Hiroya Komatsu, Sosuke Kobayashi, Naoaki Okazaki, and Kentaro Inui. 2016. Tohoku at SemEval-2016 task 6: feature-based model versus convolutional neural network for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 401–407.

Myungha Jang and James Allan. 2016. Improving automated controversy detection on the Web. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*. 865–868.

Myungha Jang and James Allan. 2018. Explaining controversy on social media via stance summarization. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*.

Myungha Jang, John Foley, Shiri Dori-Hacohen, and James Allan. 2016. Probabilistic approaches to controversy detection. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 2069–2072.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent Twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. 151–160.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: a survey. *ACM Computing Surveys (CSUR)* 50, 5 (2017), 73.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* (2016).

Anand Konjengbam, Subrata Ghosh, Nagendra Kumar, and Manish Singh. 2018. Debate stance classification using word embeddings. In *International Conference on Big Data Analytics and Knowledge Discovery (DaWaK)*. 382–395.

Michal Konkol. 2014. Brainy: a machine learning library. In *International Conference on Artificial Intelligence and Soft Computing*. 490–499.

Peter Krejzl and Josef Steinberger. 2016. UWB at SemEval-2016 task 6: stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 408–412.

Dilek Küçük. 2017a. Joint named entity recognition and stance detection in tweets. *arXiv preprint arXiv:1707.09611* (2017).

Dilek Küçük. 2017b. Stance detection in Turkish tweets. In *Proceedings of the International Workshop on Social Media World Sensors*.

Dilek Küçük and Fazli Can. 2018. Stance detection on tweets: an SVM-based approach. *arXiv preprint arXiv:1803.08910* (2018).

Kostiantyn Kucher, Carita Paradis, and Andreas Kerren. 2018. Visual analysis of sentiment and stance in social media texts. In *Proceedings of the 20th EG/VGTC Conference on Visualization (EuroVis)*.

Kostiantyn Kucher, Carita Paradis, Magnus Sahlgren, and Andreas Kerren. 2017. Active learning and visual analytics for stance classification with ALVA. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 3 (2017), 14.

Mirko Lai, Alessandra Teresa Cignarella, and Delia Irazú Hernández Farías. 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)*.

Mirko Lai, Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. 2016. Friends and enemies of Clinton and Trump: using context for detecting stance in political tweets. In *Proceedings of the Mexican International Conference on Artificial Intelligence (MICAI)*. 155–168.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance evolution and Twitter interactions in an Italian political debate. In *Proceedings of 23rd International Conference on Natural Language and Information Systems*.

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

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*. 1188–1196.

Gina-Anne Levow, Valerie Freeman, Alena Hrynkevich, Mari Ostendorf, Richard Wright, Julian Chan, Yi Luan, and Trang Tran. 2014. Recognition of stance strength and polarity in spontaneous speech. In *IEEE Spoken Language Technology Workshop (SLT)*. 236–241.

Chang Li, Aldo Porco, and Dan Goldwasser. 2018. Structured representation learning for online debate stance prediction. In *Proceedings of the International Conference on Computational Linguistics*. 3728–3739.

Junjie Lin, Wenji Mao, and Yuhao Zhang. 2017. An enhanced topic modeling approach to multiple stance identification. In *Proceedings of the ACM on Conference on Information and Knowledge Management*. 2167–2170.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on?: identifying perspectives at the document and sentence levels. In *Proceedings of the 10th Conference on Computational Natural Language Learning*. 109–116.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: state of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)* 16, 2 (2016), 10.

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing* 2 (2010), 627–666.

Can Liu, Wen Li, Bradford Demarest, Yue Chen, Sara Couture, Daniel Dakota, Nikita Haduong, Noah Kaufman, Andrew Lamont, Manan Pancholi, Kenneth Steimel, and Sandra Kübler. 2016b. IUCL at SemEval-2016 task 6: an ensemble model for stance detection in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 394–400.

Liran Liu, Shi Feng, Daling Wang, and Yifei Zhang. 2016a. An empirical study on Chinese microblog stance detection using supervised and semi-supervised machine learning methods. In *Natural Language Understanding and Intelligent Applications*. Springer, 753–765.

Nikita Lozhnikov, Leon Derczynski, and Manuel Mazzara. 2018. Stance prediction for Russian: data and analysis. *arXiv preprint arXiv:1809.01574* (2018).

Michal Lukasik, Kalina Bontcheva, Trevor Cohn, Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2019. Gaussian processes for rumour stance classification in social media. *ACM Transactions on Information Systems (TOIS)* 37, 2 (2019), 20.

Razan Masood and Ahmet Aker. 2018. The fake news challenge: stance detection using traditional machine learning approaches. In *Proceedings of the 10th International Conference on Knowledge Management and Information Sharing*.

Borislav Mavrin. 2017. *Statistical modeling of stance detection.* Ph.D. Dissertation. University of Alberta.

Qingliang Miao, Qiudan Li, and Ruwei Dai. 2009. AMAZING: a sentiment mining and retrieval system. *Expert Systems with Applications* 36, 3 (2009), 7192–7198.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

Amita Misra, Brian Ecker, Theodore Handleman, Nicolas Hahn, and Marilyn Walker. 2016. NLDS-UCSC at SemEval-2016 task 6: a semi-supervised approach to detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 420–427.

Amita Misra and Marilyn A Walker. 2013. Topic independent identification of agreement and disagreement in social media dialogue. In *Proceedings of the Conference of the Special Interest Group on Discourse and Dialogue*. 41–50.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1643–1654.

Saif M Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A dataset for detecting stance in tweets. In *Proceedings of the Language Resources and Evaluation Conference*. 3945–3952.

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

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology* 17, 3 (2017), Article 26.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of NAACL-HLT*. 767–776.

Akiko Murakami and Rudy Raymond. 2010. Support or oppose?: classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of the International Conference on Computational Linguistics*. 869–875.

Shanmugavelayutham Muthukrishnan. 2005. Data streams: algorithms and applications. *Foundations and Trends® in Theoretical Computer Science* 1, 2 (2005), 117–236.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2019. Stance classification for rumour analysis in Twitter: exploiting affective information and conversation structure. *arXiv preprint arXiv:1901.01911* (2019).

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2 (2008), 1–135.

Eli Pariser. 2011. *The filter bubble: how the new personalized web is changing what we read and how we think.* The Penguin Press, New York.

Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. 2016. JU_NLP at SemEval-2016 task 6: detecting stance in tweets using support vector machines. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 440–444.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12, Oct (2011), 2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

Rosalind W. Picard. 1997. *Affective computing.* MIT Press, Cambridge, MA, USA.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. 486–495.

Ana-Maria Popescu and Marco Pennacchiotti. 2010. Detecting controversial events from Twitter. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 1873–1876.

Mauridhi Hery Purnomo, Surya Sumpeno, Esther Irawati Setiawan, and Diana Purwitasari. 2017. Biomedical engineering research in the social network analysis era: stance classification for analysis of hoax medical news in social media. *Procedia Computer Science* 116 (2017), 3–9.

Minghui Qiu, Yanchuan Sim, Noah A Smith, and Jing Jiang. 2015. Modeling user arguments, interactions, and attributes for stance prediction in online debate forums. In *Proceedings of the SIAM International Conference on Data Mining*. 855–863.

Ashwin Rajadesingan and Huan Liu. 2014. Identifying users with opposing opinions in Twitter debates. In *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. 153–160.

Gayathri Rajendran, Bhadrachalam Chitturi, and Prabaharan Poornachandran. 2018. Stance-in-depth deep neural approach to stance classification. *Procedia Computer Science* 132 (2018), 1646–1653.

Geetanjali Rakshit, Kevin K Bowden, Lena Reed, Amita Misra, and Marilyn Walker. 2019. Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents*. Springer, 45–52.

Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems* 89 (2015), 14–46.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1650–1659.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*.

Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016. Context-sensitive Twitter sentiment classification using neural network. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*.

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

Sakala Venkata Krishna Rohit and Navjyoti Singh. 2018. Analysis of speeches in Indian parliamentary debates. *arXiv preprint arXiv:1808.06834* (2018).

Sebastian Ruder, John Glover, Afshin Mehrabani, and Parsa Ghaffari. 2018. 360° stance detection. In *Proceedings of NAACL-HLT 2018: System Demonstrations*.

Bertrand Russell. 1992. *The philosophy of Leibniz*. Routledge.

Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. 2018. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining* 8, 1 (2018), 28.

Akira Sasaki, Kazuaki Hanawa, Naoaki Okazaki, and Kentaro Inui. 2017. Other topics you may also agree or disagree: modeling inter-topic preferences using tweets and matrix factorization. *arXiv preprint arXiv:1704.07986* (2017).

Akira Sasaki, Kazuaki Hanawa, Naoaki Okazaki, and Kentaro Inui. 2018. Predicting stances from social media posts using factorization machines. In *Proceedings of the International Conference on Computational Linguistics*. 3381–3390.

Akira Sasaki, Junta Mizuno, Naoaki Okazaki, and Kentaro Inui. 2016. Stance classification by recognizing related events about targets. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*. 582–587.

Kim Schouten and Flavius Frasincar. 2016. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge & Data Engineering* 28, 3 (2016), 813–830.

Anirban Sen, Manjira Sinha, Sandya Mannarswamy, and Shourya Roy. 2018. Stance classification of multi-perspective consumer health information. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*. 273–281.

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

Gourav G Shenoy, Erika H Dsouza, and Sandra Kübler. 2017. Performing stance setection on Twitter data using computational linguistics techniques. *arXiv preprint arXiv:1703.02019* (2017).

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2018. Stance detection on microblog focusing on syntactic tree representation. In *International Conference on Data Mining and Big Data*. 478–490.

Vasiliki Simaki, Carita Paradis, and Andreas Kerren. 2017a. Stance classification in texts from blogs on the 2016 British referendum. In *Proceedings of the International Conference on Speech and Computer*. 700–709.

Vasiliki Simaki, Carita Paradis, Maria Skeppstedt, Magnus Sahlgren, Kostiantyn Kucher, and Andreas Kerren. 2017b. Annotating speaker stance in discourse: the Brexit blog corpus. *Corpus Linguistics and Linguistic Theory* (2017).

Vasiliki Simaki, Panagiotis Simakis, Carita Paradis, and Andreas Kerren. 2018. Detection of stance-related characteristics in social media text. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*.

Maria Skeppstedt, Magnus Sahlgren, Carita Paradis, and Andreas Kerren. 2016. Active learning for detection of stance components. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*. 50–59.

Maria Skeppstedt, Vasiliki Simaki, Carita Paradis, and Andreas Kerren. 2017. Detection of stance and sentiment modifiers in political blogs. In *Proceedings of the International Conference on Speech and Computer*. 302–311.

Parinaz Sobhani. 2017. *Stance detection and analysis in social media.* Ph.D. Dissertation. Université d'Ottawa/University of Ottawa.

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

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*. 551–557.

Parinaz Sobhani, Saif Mohammad, and Svetlana Kiritchenko. 2016. Detecting stance in tweets and analyzing its interaction with sentiment. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. 159–169.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. 226–234.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. 116–124.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*. 116–125.

Dhanya Sridhar, Lise Getoor, and Marilyn Walker. 2014. Collective stance classification of posts in online debate forums. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*. 109–117.

Qingying Sun, Zhongqing Wang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2018a. Stance detection via sentiment information and neural network model. *Frontiers of Computer Science* (2018). https://doi.org/10.1007/s11704-018-7150-9

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2016. Exploring various linguistic features for stance detection. In *Natural Language Understanding and Intelligent Applications*. Springer, 840–847.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018b. Stance detection with hierarchical attention network. In *Proceedings of the International Conference on Computational Linguistics*. 2399–2409.

Sahil Swami, Ankush Khandelwal, Manish Shrivastava, and S Sarfaraz-Akhtar. 2017. LTRC_IIITH at IberEval 2017: stance and gender detection in tweets on catalan independence. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*.

Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. An English-Hindi code-mixed corpus: stance annotation and baseline system. *arXiv preprint arXiv:1805.11868* (2018).

Mariona Taulé, M Antonia Martí, Francisco Rangel, Paolo Rosso, Cristina Bosco, and Viviana Patti. 2017. Overview of the task on stance and gender detection in tweets on Catalan independence at IberEval 2017. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*.

Theano Development Team. 2016. Theano: a Python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688* (2016).

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 327–335.

James Thorne, Mingjie Chen, Giorgos Myrianthous, Jiashu Pu, Xiaoxuan Wang, and Andreas Vlachos. 2017. Fake news stance detection using stacked ensemble of classifiers. In *Proceedings of the EMNLP Workshop: Natural Language Processing Meets Journalism*. 80–83.

Benjamin Timmermans, Tobias Kuhn, Kaspar Beelen, and Lora Aroyo. 2017. Computational controversy. In *Proceedings of the International Conference on Social Informatics*. 288–300.

Adam Tsakalidis, Nikolaos Aletras, Alexandra I Cristea, and Maria Liakata. 2018. Nowcasting the stance of social media users in a sudden vote: the case of the Greek referendum. In *Proceedings of the ACM International Conference on Information and Knowledge Management*.

Martin Tutek, Ivan Sekulic, Paula Gombar, Ivan Paljak, Filip Culinovic, Filip Boltuzic, Mladen Karan, Domagoj Alagić, and Jan Šnajder. 2016. Takelab at SemEval-2016 task 6: stance classification in tweets using a genetic algorithm based ensemble. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 464–468.

C. J. van Rijsbergen. 1979. *Information retrieval* (2d ed. ed.). Butterworths London; Boston.

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

Prashanth Vijayaraghavan, Ivan Sysoev, Soroush Vosoughi, and Deb Roy. 2016. DeepStance at SemEval-2016 task 6: detecting stance in tweets using character and word-level CNNs. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*.

R. Vinayakumar, S. Sachin Kumar, B. Premjith, Poornachandran Prabaharan, and Kotti Padannayil Soman. 2017. Deep stance and gender detection in tweets on Catalan independence@IberEval 2017. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: task definition and dataset construction. In *Proceedings of the ACL Workshop on Language Technologies and Computational Social Science*. 18–22.

Marilyn A Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012a. Stance classification using dialogic properties of persuasion. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 592–596.

Marilyn A Walker, Pranav Anand, Rob Abbott, Jean E Fox Tree, Craig Martell, and Joseph King. 2012b. That is your evidence?: classifying stance in online political debate. *Decision Support Systems* 53, 4 (2012), 719–729.

Byron C Wallace. 2015. Computational irony: a survey and new perspectives. *Artificial Intelligence Review* 43, 4 (2015), 467–483.

Penghui Wei, Junjie Lin, and Wenji Mao. 2018a. Multi-target stance detection via a dynamic memory-augmented network. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1229–1232.

Penghui Wei, Wenji Mao, and Daniel Zeng. 2018b. A target-guided neural memory model for stance detection in Twitter. In *Proceedings of the International Joint Conference on Neural Networks*. 1–8.

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at SemEval-2016 task 6: a specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 384–388.

Michael Wojatzki and Torsten Zesch. 2016a. ltl.uni-due at SemEval-2016 task 6: stance detection in social media using stacked classifiers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 428–433.

Michael Wojatzki and Torsten Zesch. 2016b. Stance-based argument mining–modeling implicit argumentation using stance. In *Proceedings of the KONVENS Conference*. 313–322.

Michael Wojatzki and Torsten Zesch. 2017. Neural, non-neural and hybrid stance detection in tweets on Catalan independence. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*.

Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. 2016. Quantifying political leaning from tweets, retweets, and retweeters. *IEEE Transactions on Knowledge and Data Engineering* 28, 8 (2016), 2158–2172.

Chang Xu, Cecile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. *arXiv preprint arXiv:1805.06593* (2018).

Jiaming Xu, Suncong Zheng, Jing Shi, Yiqun Yao, and Bo Xu. 2016a. Ensemble of feature sets and classification methods for stance detection. In *Natural Language Understanding and Intelligent Applications*. Springer, 679–688.

Kang Xu, Sheng Bi, and Guilin Qi. 2017. Semi-supervised stance-topic model for stance classification on social media. In *Proceedings of Joint International Semantic Technology Conference*. 199–214.

Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016b. Overview of NLPCC shared task 4: stance detection in Chinese microblogs. In *Natural Language Understanding and Intelligent Applications*. Springer, 907–916.

Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. 2019. Unsupervised fake news detection on social media: a generative approach. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*.

Tae Yano, Philip Resnik, and Noah A Smith. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL-HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. 152–158.

Nan Yu, Da Pan, Meishan Zhang, and Guohong Fu. 2016. Stance detection in Chinese microblogs with neural networks. In *Natural Language Understanding and Intelligent Applications*. Springer, 893–900.

Guido Zarrella and Amy Marsh. 2016. MITRE at SemEval-2016 task 6: transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 458–463.

Qiang Zhang, Emine Yilmaz, and Shangsong Liang. 2018. Ranking-based method for news stance detection. In *Proceedings of the Web Conference Companion*.

Shaodian Zhang, Lin Qiu, Frank Chen, Weinan Zhang, Yong Yu, and Noémie Elhadad. 2017. We make choices we think are going to save us: debate and stance identification for online breast cancer CAM discussions. In *Proceedings of the International Conference on World Wide Web Companion*. 1073–1081.

Zhihua Zhang and Man Lan. 2016. ECNU at SemEval 2016 task 6: relevant or not? supportive or not? a two-step learning system for automatic detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 451–457.

Yiwei Zhou, Alexandra I Cristea, and Lei Shi. 2017. Connecting targets to tweets: semantic attention-based model for target-specific stance detection. In *Proceedings of the International Conference on Web Information Systems Engineering*. 18–32.

Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*. 307–314.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018a. Detection and resolution of rumours in social media: a survey. *ACM Computing Surveys (CSUR)* 51, 2 (2018), 32.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. In *Proceedings of the International Conference on Computational Linguistics*. 2438–2448.

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

A SUPPLEMENTARY MATERIAL

Two subsections of the paper (titled "Some Remarks on Approaches to Stance Detection" and "Observations and Recommendations for Stance Detection Researchers", respectively) are provided as online supplementary material.