Review

Robin Schaefer* and Manfred Stede

Argument Mining on Twitter: A survey

https://doi.org/10.1515/itit-2020-0053
Received December 4, 2020; revised February 11, 2021; accepted March 15, 2021

Abstract: In the last decade, the field of argument mining has grown notably. However, only relatively few studies have investigated argumentation in social media and specifically on Twitter. Here, we provide the, to our knowledge, first critical in-depth survey of the state of the art in tweet-based argument mining. We discuss approaches to modelling the structure of arguments in the context of tweet corpus annotation, and we review current progress in the task of detecting argument components and their relations in tweets. We also survey the intersection of argument mining and stance detection, before we conclude with an outlook.

Keywords: Argument Mining, Twitter, Stance Detection

ACM CCS: Computing methodologies → Artificial intelligence → Natural language processing → Information extraction

1 Introduction

In recent years, the discipline of argument mining (AM), which concentrates on the intersection of computational linguistics and computational argumentation, has grown notably [28]. However, while the majority of research focuses on well-structured text genres like persuasive essays [27], legal texts [20] or Wikipedia articles [17], a comparatively small amount of work exists on AM in social media [13]. In particular, the microblogging service Twitter[1] appears to be a source of argumentative texts which has received only little attention [6, 1, 25].

Given the increased usage of Twitter in political online discourse, investigating the extraction of argumentative text from tweets becomes especially important. However, previous research has shown that conducting AM on Twitter is a challenging task, which originates in part from the relatively high degree of noisy and irregular language used in this kind of data [30]. Also, strict licensing conditions complicate the distribution of annotated tweet corpora.[2]

While these issues have hindered progress on the task, it is fair to say that previous work also proved its feasibility. Crucially, work on tweet-based AM not only provides tools for extracting and analysing a crucial sub-group of argumentative texts. It is also of interest for the broader AM community, as innovative approaches are tested during the development of Twitter-specific AM systems.

In this paper we provide the, to our knowledge, first critical in-depth survey of the state of the art in tweet-based AM. In particular, we focus on three tasks: 1) corpus annotation, 2) argument component and relation detection, and 3) stance detection.

Corpus annotation

Corpus annotation can represent a severe bottleneck for progress in AM, like in many research areas in computational linguistics, because successful model training usually depends on well-annotated data. So far only a few studies have focused on the creation of tweet corpora for argumentation. However, important steps were completed to develop argument modelling approaches suitable for tweets.

Argument and relation detection

First work on detecting arguments in tweets has primarily focused on identifying argument components on the full tweet level. More recent work has started to explore the detection of *Argumentative Discourse Units* (ADU) [22] within tweets. Only little work exists on relation detection. However, solving this task is a necessary precondition for tasks like argument graph building.

Stance detection

Stance detection and AM are closely interrelated given that both disciplines revolve around extracting standpoints to-

[1] https://twitter.com/

[2] From an ethical viewpoint we consider these restrictions reasonable, as we are dealing with user-generated data.

***Corresponding author: Robin Schaefer,** University of Potsdam, Dept. Linguistics, D-14476 Potsdam, Germany, e-mail: robin.schaefer@uni-potsdam.de, ORCID: https://orcid.org/0000-0003-2904-3410
Manfred Stede, University of Potsdam, Dept. Linguistics, D-14476 Potsdam, Germany, e-mail: stede@uni-potsdam.de

wards a given topic. In addition, AM is also interested in constellations of reasons and counter-considerations for a standpoint. Given the relatedness of both disciplines, some researchers have adopted joint approaches to AM and stance detection.

As different studies are discussed in this paper, inevitably variant terminology and notation arise. Primarily, we will refer to notation used in [25] and partly in [1]. In particular, we use the terms *claim* and *evidence* for the two core components of an argument: a claim refers to a standpoint to a topic, whereas evidence is defined as supportive or opposing information with respect to the claim. In addition, we use ADU as a generic term for claim and evidence units.

Table 1: Examples 1–4 of Argumentative Tweets from the Literature.

	Tweet Examples
(1)	#climateprotection is one of the most important topics, but we won't solve it with bans. Not many want to give up progress, previous generations achieved with innovation. [translated from German] [25]
(2)	Please who don't understand encryption or technology should not be allow to legislate it. There should be a test... https://t.co/I5zkvK9sZf [1]
(3)	Perhaps Apple can start an organ harvesting program. Because I only need one kidney, right? #iPadPro #AppleTV #AppleWatch [5]
(4)	Rioters in Birmingham make moves for a CHILDREN's hospital, are people that low? #Birminghamriots [23]

Table 1 shows tweet examples with typical issues related to tweet-based argumentation.[3] While the first example could be annotated as containing a claim with supportive evidence, example (2) is more difficult to assess due to the usage of a link as evidence. Examples (3) and (4) both contain rhetorical questions, which tend to be a frequent instrument to express one's opinion on Twitter but may pose a challenge for AM systems. In addition, example (3) is clearly sarcastic, which increases difficulties as well. Thus, the examples show that identifying argumentation in tweets is far from trivial and AM researchers need to decide how to deal with peculiarities typical for user-generated data.

This paper is structured as follows: in Section 2 we describe the motivation for conducting AM on Twitter in more detail. In Section 3 we discuss current approaches to argument modelling for corpus annotation and the actual process of creating an annotated argument tweet corpus. Section 4 focuses on practical approaches to detect argument components and their relations in tweets, before we survey the intersection of AM and stance detection in Section 5. We conclude in Section 6 with an outlook on possible future developments of AM on Twitter.

2 Motivating AM on Twitter

Given its fast-paced and sometimes superficial nature, it is reasonable to question if argumentation actually takes place on Twitter. We will therefore motivate the endeavor of investigating tweet-based AM, before we continue with the critical survey of the field's state of the art. We undertake this by focusing on three questions:

1. Does Twitter cover topics that allow for diverse standpoints?
2. Are these topics argumentatively discussed?
3. Which adjustments to already existing AM approaches follow from Twitter-specific characteristics?

Question 1 is based on the hypothesis that topics provoking diverse opinions are more likely to actually trigger argumentation and can be answered straightforwardly. We take this to be the case, as indeed, many controversial topics have been covered on Twitter in recent years, including, for instance, climate change [29] and abortion rights [32].

Given that controversial topics are covered on Twitter, we need to ask if these are argumentatively discussed (Question 2). To our knowledge, no statistics on the amount of tweet-based argumentation exist. However, previous research investigated the nature of Twitter conversations, e. g. by concentrating on co-reference resolution across tweet boundaries [3]. While a conversation does not need to be argumentative per se, we consider it fair to conclude that a conversation about a controversial topic likely results in argumentation. This assumption depends, however, on how argumentation is defined. Often, an argument is understood to consist of a claim and at least one evidence unit. If we accept this definition a substantial number of opinions on Twitter would indeed be non-argumentative due to missing evidence. As a microblogging service Twitter imposes a strict constraint on the allowed tweet length (280 characters). Thus, there are technical restrictions imposed that hinder the formulation of more complex argumentative structures. For our purposes, we treat a single claim also as an argumentative unit which

[3] Note that Example (1) is part of the corpus presented in [25], but was not explicitly mentioned in the paper.

should be considered by an AM system (similar to [5, 6]). However, the quality of such an argumentative portion remains open for debate.

Using Twitter as a data source for AM entails certain constraints that shape the development of new approaches (Question 3). As mentioned, tweets tend to contain a substantial amount of noisy data, which include spelling and grammar mistakes and innovations. Further, a number of social media/Twitter conventions are frequently applied, for example hashtags, emoticons, abbreviations, etc. [30]. These may be subsumed under the label *linguistic characteristics*. In addition, Twitter has properties that require researchers to perform adaptations to this new data type. In principle, Twitter allows for two different ways to post original[4] tweets.[5] 1) One can independently post a tweet. A common strategy is the use of hashtags to link the tweet to a topic. 2) One can make use of Twitter's reply functionality. This links one's own tweet to a previously published tweet. A series of tweets in a reply relation is called *conversation*. Importantly, both strategies may lead to different kinds of argumentation. A Twitter conversation resembles a dialogue and argumentation will reflect this. For instance, users may attack each others' claims directly or reject other opinions altogether. This dialogic context is generally missing if tweets are posted independently. Still, we see these tweets as potentially argumentative given that they may contribute to the discussion. In Section 3 we will show that these differences in posting behavior have implications for the way arguments are modelled.

All papers discussed in more detail are listed in Table 2. The abbreviations are used throughout this paper for easier reference. Importantly, we focus on practical approaches to tweet-based AM which tackle it primarily from a machine learning viewpoint. Other research, like the formal modelling approach of [14], is beyond the scope of this paper.

3 Creating annotated tweet corpora

Until today only a few studies have been conducted on argument annotation in tweets, hence the small amount of annotated corpora suitable for tweet-based AM. As the training of AM systems heavily depends on annotated data, this represents a major limitation to the progress in the field. However, important first work on argument annotation has been conducted, including Twitter-related adjustments to the task. Corpora and inter-annotator agreement scores (IAA)[6] discussed in this section are listed in Table 3. The used annotation schemes are described in Table 5.

Table 2: Papers on AM and Stance Detection on Twitter (CA=corpus annotation, AD=argument or relation detection, SD=stance detection).

Authors	Abbr.	Task
Addawood & Bashir (2016)	AB2016	CA, AD
Addawood et al. (2017)	ASB2017	CA, SD
Bosc et al. (2016) a	BCV2016a	CA
Bosc et al. (2016) b	BCV2016b	AD
Dusmanu et al. (2017)	DCV2017	CA, AD
Llewellyn et al. (2014)	LGOK2014	AD
Procter et al. (2013)	PVV2013	CA
Schaefer & Stede (2019)	SS2019	SD
Schaefer & Stede (2020)	SS2020	CA, AD
Wojatzki & Zesch (2016)	WZ2016	CA, SD

Table 3: Corpus Annotation Results (a=argument, c=claim, e=evidence, r=relation, s=stance, f=factual, sc=source, multi=a vs c vs e; cϰ=Cohen's ϰ, d=Dice, fϰ=Fleiss' ϰ, α=Krippendorff's α, scϰ = Siegel and Castellan's ϰ).

Paper	Corpus Size	Class	IAA
Unit of Annotation: Tweet			
AB2016	3,000	a	0.67 (cϰ)
		e	0.79 (cϰ)
BCV2016a	4,000	a	0.74 (α)
		r	0.67 (α)
DCV2017	987+900	a	0.77 (scϰ)
		f	0.73 (scϰ)
		sc	0.84 (d)
PVV2013	7,729	c, e	0.89–0.96 (–)
SS2020	300	a	0.53 (cϰ)
		c	0.55 (cϰ)
		e	0.44 (cϰ)
WZ2016	733	s	0.63 (fϰ)
Unit of Annotation: ADU			
SS2020	300	multi	0.38 (cϰ)
		a	0.45 (cϰ)

Initial attempts on annotating tweets for argumentation were conducted by [23] (henceforth PVV2013). Contrasting with following work, PVV2013 were not specifically interested in advancing AM, but instead focused on

[4] Crucially, this excludes the significant group of retweets.

[5] Importantly, for now we ignore a third way: a thread. A thread is a series of tweets posted by one user. To our knowledge, it has not been studied in the context of AM.

[6] For a discussion on IAA see [4].

Table 4: Examples 5-8 of Argumentative Tweets from the Literature.

	Paper	Tweet Examples
(5)	PVV2013	Girlfriend has just called her ward in Birmingham Children's Hospital & there's no sign of any trouble #Birminghamriots
(6)	BCV2016a	[Tweet A] The letter #47Traitors sent to Iran is one of the most plainly stupid things a group of senators has ever done. http://t.co/oEJFlJeXjy [Tweet B] Republicans Admit: That Iran Letter Was a Dumb Idea http://t.co/Edj57f4nE8. You think?? #47Traitors
(7)	AB2016	I care about #encryption and you should too. Learn more about how it works from Mozilla at https://t.co/RTFiuTQXyQ
(8)	SS2020	[Context Tweet] Liberals who think climate protection is equal to deprivation of liberty confound liberty and selfishness. [Reply Tweet][Example (1) in Table 1] #climateprotection is one of the most important topics, but we won't solve it with bans. Not many want to give up progress, previous generations achieved with innovation.

analysing the flow of information on Twitter during a crisis situation. The corpus contains English tweets related to the London Riots of 2011 (see Example (5) in Table 4).

Concentrating on the full tweet level, PVV2013 developed an annotation scheme that first required the grouping of tweets according to their type (e. g. media reports, rumours or reactions). Second, each type had its sub-scheme focusing more on content and function. For this paper, the *rumour* type is important, as its sub-scheme is based upon the notions of *claim*, *counterclaim* and *evidence*. While being borrowed from argumentation theory, these terms were fully understood in the context of rumours. Specifically, a claim is basically considered as a rumour. Interestingly, the authors annotated tweets for counterclaims as well, thereby taking the dialogic aspect of argumentation into account. With respect to AM on Twitter, this category tends to be ignored. Evidence is defined as supporting or challenging content relating to a (counter)claim. However, evidence provided with additional information (e. g. links to news articles) would likely be subsumed under the type *media report*. This indicates that evidence is restricted to contributing but unproven information. While making sense in the context of this study, these definitions of argument components could lead to misunderstandings. Also, they excluded argumentation unrelated to rumours, which may be one reason why this corpus remains little considered in the field of tweet-based AM (with the notable exception of [18] (see Section 4)).

Another tweet corpus annotated for argumentation was created by [5] (henceforth BCV2016a). This data set, called *Dataset of Arguments and their Relations on Twitter* (DART), was developed with the core tasks of the AM pipeline in mind, i. e. the extraction of argument components and their relations. It contains 4000 English tweets on four topics related to politics (e. g. the Grexit) and product releases (e. g. the Apple iWatch).

The authors decided on using the full tweet as the unit of annotation and applied the categories *argumentative/non-argumentative/(unknown)*, thereby refraining from differentiating further between claim and evidence. The category *argumentative* subsumed the notions of *opinion*, *claim/conclusion*, *premise* and *factual information*. While these terms appear to be based on a more intuitive conceptual understanding, [6] (henceforth BCV2016b), who used the same data and annotations in their work, gave more details on the definition of argumentativeness underlying DART. A tweet is to be annotated as *argumentative* if it contains parts of the standard argument structure, i. e. claim or evidence, where a claim is understood as a conclusion derived from evidence. This is justified due to the typically incomplete argument structures found on Twitter. The Krippendorff's α of 0.74 indicates that the expert annotators agreed on a substantial number of annotations, which might be favored by the simple binary[7] annotation task, i. e. argumentative vs non-argumentative. The actual annotation of the full corpus (4000 tweets) was conducted by three recruited students with a non-linguistic background, which had been trained on doing the task. Final annotations per tweet were derived using majority voting. Annotations of a tweet subset were compared with the annotations conducted by the already mentioned expert annotators and yielded a Krippendorff's α of 0.81.

[7] Technically, this is a three-way annotation task (argumentative vs non-argumentative vs unknown). However, as *unknown* is treated as a loophole if the other categories cannot be assigned, we see the annotation task as binary.

Table 5: Annotation Schemes.

Paper	Classes	Definitions
PVV2013	Claim	A rumour without supporting information
	Claim + Evidence	A claim + supporting/challenging information
	Counterclaim	A claim disputing another claim.
	Counterclaim + Evidence	A disputing claim + supporting/challenging information
BCV2016a	**Step I: Argumentativeness**	
	Argumentative	Containing a claim or evidence
	Non-argumentative	Not containing a claim or evidence
	Step III: Relations (Step II: pair creation)	
	Support	A positive relation between tweets
	Attack	A negative relation between tweets
DCV2017	**Step I (see BCV2016a (Step I))**	
	Step II	
	Factual	Containing proven information or reported speech
	Opinion	Not containing proven information or reported speech
AB2016	**Step I: Argumentativeness**	
	Argumentative	Containing claim (+ evidence)
	Non-argumentative	Not containing a claim
	Step II: Evidence	
	News media account	Content from news media account (often via link)
	Expert opinion	Opinion of someone having more experience (experts)
	Blog post	A link to a blog post
	Picture	A shared picture
	Other	Evidence types not included above (audio, books, etc.)
	No evidence	Not containing evidence
SS2020	Claim	A standpoint towards a topic
	Evidence (relating to reply tweet)	Evidence related to claim in reply tweet
	Evidence (relating to context tweet)	Evidence related to claim in context tweet
	Evidence (relating to both tweets)	Evidence related to claims in both tweets

In addition to the full tweet annotations, BCV2016a conducted relation annotations. For this step, the authors basically considered the full tweets as nodes in an argument graph, which can be linked via their relations. For annotating relations, argumentative tweets had to be grouped into pairs. Importantly, this pairing was based on content similarity instead of a reply relation between tweets within a given Twitter conversation. BCV2016a justified this decision by arguing that Twitter users tend to give their opinion with respect to a certain topic, e. g. by using hashtags, and less by directly replying to other users. While the authors raised a valid point, given that indeed a notable number of tweets is posted independently, we still consider both independent and reply tweets as potential sources for argumentation (see Section 2 for a discussion). To automatically group tweets into pairs, BCV2016a annotated tweets according to an additional set of semantic categories (e. g. topic: *product release*; categories: *price*, *look*, *advertisement*, etc.). Afterward, a classifier was trained on these annotations. Tweets classified into the same category were randomly grouped into pairs.

In a final step, expert annotators annotated the tweet pairs according to the categories *support*, *attack*, and *unrelated*. *Support* is defined as a positive relation between tweets, which is the case, for instance, when both tweets express similar views on the topic (see Example (6) in Table 4). In contrast, *attack* is understood as a negative relation between tweets. In parallel, tweet pairs were annotated for entailment. This rested upon the assumption that support/entailment and attack/contradiction were conceptually related, respectively. BCV2016b noted that during relation annotation the temporal dimension was taken into account, i. e. Tweet A only can support/attack Tweet B if it was posted at a later time step. While this intuitively seems reasonable, it actually conflicts with the authors assumption on the way argumentation takes place on Twitter. BCV2016a decided on relying on a tweet's semantic relatedness to a given topic (e. g. the Grexit). This, however, implies that the temporal unfolding of a discussion needs to be ignored, given that it is difficult to determine if a tweet's author actually has read a previously posted tweet.

Moreover, the assignment of tweet pairs within a category was conducted randomly, which placed a strong weight on the category annotations. As the annotation results show that a vast majority of tweet pairs was assigned the *unrelated* label, one may argue that the categories were not adequately fine-grained for the task. Also, the authors pointed out that some tweets within a pair are too complex to be consistently annotated with one label. This issue could be approached by allowing relations to be partial, as proposed by BCV2016a. Alternatively, this could be solved by additionally annotating spans within tweets, as conducted by [25] (henceforth SS2020).

In subsequent work, [11] (henceforth DCV2017) concentrated on the task of source identification. They extended the annotations of the #Grexit subset (size: 987) of DART with two additional layers: 1) factual vs opinion, and 2) source. A factual tweet needed to contain provable information or reported speech. An opinionated tweet, in contrast, did not contain any of these contents. All factual tweets were further annotated for the source, if any source existed. Parallel to the annotations of the #Grexit data, DCV2017 additionally conducted the same annotations for 900 tweets on #Brexit, after having them labelled as argumentative/non-argumentative (Step 1 in BCV2016a).

In a similar vein, another important approach to annotating argument structure in tweets was proposed by [1] (henceforth AB2016). Contrasting with BCV2016a, argument relations were ignored. Instead the authors placed a strong focus on evidence types. The annotated corpus contains 3000 English tweets on the encryption debate of 2016 concerning Apple and the FBI. Similar to BCV2016a, tweets were collected independently using a keyword (“encryption”) (see Example (7) in Table 4). No reply relation information was taken into account.

Annotations were conducted using a two-step approach. First, two annotators labelled a given tweet as *argumentative* or *non-argumentative*. Second, the same annotators further annotated argumentative tweets according to a set of different evidence types. Annotators focused on the full tweet as the unit of annotation.

AB2016 considered a tweet as argumentative if it contains an opinion, i. e. a claim, and optional evidence for the opinion. Tweets containing no argumentative content or potential evidence without an opinion are treated as *non-argumentative*. The notion of argumentativeness is based to a certain degree on an intuitive understanding, given that opinion/claim remains undefined. While evidence types are precisely defined, annotators have to decide if the no-evidence content is sufficiently opinionated to be rated as argumentative. Granted, the Cohen’s κ score of 0.67 indicates that annotators were indeed able to perform this task, however, cross-study comparisons are hindered.

This work’s innovation lies in the detailed classification of evidence types, which draws its strength from the consideration of types frequently used in social media. While this comprises news media accounts, blog posts and expert opinions, also non-linguistic data like pictures are included. The latter is usually not part of the AM field, which focuses on extracting argumentation from language resources. However, if we consider Twitter as a data source per se, including pictures may result in a more accurate representation of used argumentative structures and strategies. Links represent another non-linguistic category, which is commonly used to provide evidence via external resources. Due to Twitter’s severe constraints on a tweet’s complexity by imposing a strict character limit, we consider it appropriate to accept links as evidence. Still, the question remains to be answered how such linked evidence is incorporated. For instance, it is possible to use evidence links as binary tweet labels (e. g. *contains evidence*/*contains no evidence*). Alternatively, it may be worth considering including the actual content of the external resource, although this likely increases the task’s complexity. Also, technical and license-related constraints needed to be considered. By placing a focus on evidence types, this work partially relates to PVV2013. However, AB2016 use *evidence* as a cover term for different types, whereas PVV2013 have a more narrow understanding of *evidence* and treat verifiable evidence differently.

More recent work by SS2020 investigated the feasibility of annotating ADU spans within tweets, thereby presenting the first approach to argumentation annotation on Twitter not explicitly limited to the full tweet level. The annotated corpus contains 300 German tweets on the climate change debate from 2019. Tweets were collected using the keyword *klima* (“climate”).

Another difference to previous work is associated with the corpus structure. It is based on pairs consisting of tweets in a reply relation instead of independent tweets. Hence, some conversation information, although being somewhat limited, is used. The first tweet, called *context tweet*, provides context information which is assumed to have a facilitating effect on the annotation procedure. The second tweet is a direct reply to the context tweet, hence called *reply tweet* (see Example (8) in Table 4). Only the reply tweet was annotated in this study.

The annotation scheme is based on the two core components of argumentation: claim and evidence. SS2020 defined a claim as a standpoint towards a topic (e. g. climate change). In contrast, evidence was defined as a supportive or objective statement relating to a claim. Thus, while a claim can potentially function as an independent unit, evidence always stands in relation to a claim and cannot appear on its own. In this regard, SS2020 defined evidence similarly to PVV2013 and AB2016. Contrasting with AB2016, however, who discriminated between different evidence categories (e. g. news or expert opinion), SS2020 separated it further into sub-categories with respect to the tweet that contains the target of the evidence, i. e. the reply tweet, the context tweet or both.

With respect to the context tweet it is important to underline its twofold function. First, it is used as a context to improve understanding of the reply tweet's content. Second, by allowing the target claim of an evidence unit to occur in the context tweet, the authors implicitly indicated that reply relations between tweets potentially play a role in the unfolding argumentation. This, however, contrasts with BCV2016a who argued that Twitter users rarely use the reply function to post argumentative content, but instead simply broadcast it by posting independent tweets. Given that we see both types of tweet posting as relevant for AM, we consider SS2020 approach as justified. Still, as in BCV2016a, not the full picture is included, as independently posted tweets are ignored. Under the assumption that argumentation may unfold differently depending on the type of posting, this issue may require contemplation in future research.

The last corpus we mention in this paper has been presented by [31] (henceforth WZ2016). Importantly, this work focused on stance annotation and will be discussed in more detail in Section 5.

Interim conclusion

We conclude that most studies focused on the core components of argumentation: claim and evidence (AB2016, PVV2013, SS2020). In contrast, relation annotation has only been rarely investigated (BCV2016a). So far, annotations were usually conducted on the full tweet level (AB2016, BCV2016a, PVV2013), with one exception (SS2020). Importantly, some researchers have taken tweet types (independent vs reply) into account and adjusted their study design accordingly (BCV2016a, SS2020).

4 Practical approaches to AM on Twitter

While corpus annotation undoubtedly plays a crucial role both for the creation of AM-suitable data sets and for argument modelling in general, the actual engineering side of AM starts after the data has been collected and prepared. We identify the following practical AM tasks:[8] 1) (general) argument detection, 2) claim detection, 3) evidence (type) detection, and 4) relation detection. Although most previous studies focused on more than one task, we decided on discussing the tasks consecutively in respective subsections. A survey of the results is given in Table 6.

Table 6: Component and Relation Detection Results (Cohen's Kappa=cϰ, P=Precision, R=Recall).

Paper	Unit	F1/(cϰ)	P	R
(General) Argument Detection				
AB2016	Tweet	0.89	0.89	0.89
BCV2016b	Tweet	0.78	–	–
DCV2017	Tweet	0.78	0.80	0.70
SS2020	Tweet	0.82	0.80	0.86
	ADU	0.72	0.73	0.72
Claim Detection				
LGOK2014	Tweet	0.79–0.86 (cϰ)	–	–
SS2020	Tweet	0.82	0.80	0.85
	ADU	0.59	0.60	0.59
Evidence Detection				
AB2016	Tweet	0.79	0.79	0.80
DCV2017	Tweet	0.80	0.81	0.79
SS2020	Tweet	0.67	0.68	0.68
	ADU	0.75	0.76	0.76
Relation Detection				
BCV2016b	Tweet	0.00–0.20	–	–

4.1 (General) argument detection

Given that presumably a substantial number of tweets is non-argumentative in nature, developing a system that can separate argumentative from non-argumentative tweets appears to be a valid starting point. Indeed, most studies concentrated on general argument detection to a certain degree. As most data sets are based on full tweet annotations, approaching this task by means of supervised classification is a common choice, although differences exist with respect to the actually used classification algorithm.

AB2016 presented argument classification experiments using Naive Bayes (NB), Support Vector Machines (SVM) and Decision Trees (DT). Best results (F1: 0.89) were

[8] Argument, claim and evidence detection refer both to classification on the full tweet level and to ADU detection within tweets.

achieved by training an SVM model on a feature set consisting of n-grams, psychometric features (e.g the presence of tokens referring to psychological processes), linguistic features (e. g. POS percentages) and Twitter-related features (e. g. the presence of hashtags). The authors considered general argument detection as a necessary first step, before they continued with the task of evidence type detection (see Section 4.3).

BCV2016b presented an implemented pipeline trained on the DART corpus. The first step of this pipeline involved the separation of argumentative tweets from non-argumentative tweets. The authors chose a Logistic Regression (LR) approach based on the following features: unigrams, bigrams, POS Tags, and bigrams of POS Tags. Using all features yielded an F1 score of 0.78. Later steps involved relation detection (see Section 4.4).

DCV2017 used a combination of the DART sub-corpus referring to the topic *#Grexit* and an own tweet set on the topic *#Brexit*, which has been annotated with the first step of the annotation scheme of BCV2016a and an additional own annotation scheme. Their work on argument detection represented the first step of an AM pipeline also involving fact recognition and source identification (see Section 4.3). The authors experimented with LR and Random Forest (RF) approaches. Models were trained on lexical (e. g. n-grams), Twitter-related (e. g. emoticons), syntactic/semantic (e. g. dependency relations) and sentiment features. An LR model trained on a combination of all features yielded best results (F1: 0.78), and replicated scores presented by BCV2016b.

Given that the corpus presented in SS2020 is based both on full tweet and ADU annotations, two different approaches on general argument detection were chosen. To the best of our knowledge, SS2020 presented the first tweet-based AM results relying on ADU annotations. In line with the previously presented studies classification models were trained on full tweet annotations. The authors decided on using an eXtreme Gradient Boosting (XGBoost) model [8] that was trained on different sets of n-grams and pretrained BERT [10] document embeddings, respectively. The model trained on the latter yielded better results (F1: 0.82). To perform argument unit detection, the authors decided on applying a sequence labeling approach based on Conditional Random Fields (CRF) [16]. Three types of features were used: 1) unigrams, 2) a set of linguistic (e. g. POS Tags) and Twitter-related (e. g. hashtags) features and 3) pretrained BERT word embeddings. The model trained on BERT embeddings yielded best results (F1: 0.72). In the same vein, SS2020 also conducted claim and evidence detection (see Sections 4.2 and 4.3).

To summarize, current approaches to general argument detection are usually based on supervised classification. The studies show that SVM, LR and XGBoost models yielded best results, respectively. If ADU spans are considered instead, sequence labeling becomes a more suitable approach. Typically used feature sets include lexical, linguistic or Twitter-related features. More recent approaches are also based on BERT embeddings. We consider general argument detection as an initial step that could represent an important filter for downstream tasks in an AM pipeline. However, more fine-grained component detection is needed if more detailed argument structures are to be extracted.

4.2 Claim detection

We define claim detection as the task of identifying opinions and standpoints with respect to a topic. It may be applied to a data set already pre-filtered by general argument detection or, depending on the use case, as an initial step itself, thereby replacing the argument detection step. Further, it may be approached in conjunction with evidence detection or as a complementary but independent task. As in (general) argument detection, existing approaches to claim detection are heavily based on supervised classification and sequence labeling.

Early results on tweet-based claim (and joint evidence) detection were presented by [18] (henceforth LGOK2014). This study was based on the annotated *rumour* sub-corpus created by PVV2013, which contains annotations for the classes *claim*, *claim with evidence*, *counterclaim* and *counterclaim with evidence*. Thus, evidence is only considered in combination with a claim and cannot be detected independently. Also, counterclaims are introduced as a separate class, which contrasts with other studies on tweet-based AM. LGOK2014 experimented with different feature sets (e. g. unigrams, bigrams of POS tags, punctuation) and classification algorithms (NB, SVM, DT). A DT model performed most reliably for the different sub-tasks by using unigrams (F1: 0.79–0.86). Additional feature analyses showed that results can be improved by augmenting the unigram feature set with other feature types, out of which bigrams of POS tags were the most promising.

SS2020 conducted claim detection in parallel to general argument detection, i. e. classification and sequence labeling were used for full tweet and ADU annotations, respectively. As features again n-grams and pretrained BERT document embeddings were chosen for the classification task. The sequence labeling approach was based on unigrams, a set of linguistic and Twitter-related features and pretrained BERT word embeddings. The same ML models were chosen, i. e. XGBoost for classification and CRF for sequence labeling. However, while the classification results (F1: 0.82) were similar to those obtained during argument detection, sequence labeling results (F1: 0.59) were comparatively low.

To summarize, successful implementations of claim detection approaches exist. In the respective studies, DT and XGBoost yielded the most promising classification results. CRF were applied for sequence labeling. Moreover, best results are based on unigram features augmented with more sophisticated feature types or on BERT embeddings.

4.3 Evidence detection

We define evidence detection as the task of identifying evidence statements given with respect to a certain claim. While evidence does not exclusively include factual, i. e. provable, statements, previous work also focused on fact recognition. In this paper we understand the latter as a sub-task of evidence detection and, hence, discuss it in this sub-section.

AB2016 heavily based their AM approach on evidence type detection. After having identified argumentative tweets in a first filter step, concrete evidence types were detected afterwards. The authors used NB, SVM and DT models and trained them on sets of different features (n-grams, psychometric, linguistic and Twitter-related features). An SVM trained on all features yielded best results (F1: 0.79). In addition, one-vs-all classification results were calculated for the three largest evidence type classes, i. e. *News*, *Blog* and *No evidence* and per feature set. F1 averages across evidence types indicated that basic n-gram features already performed well (F1: 0.80), whereas using all features increased scores only slightly (F1: 0.83). Interestingly, using independent sets of psychometric, linguistic or Twitter-related features yielded lower results.

DCV2017 approached evidence detection via fact recognition. The authors considered the task as especially relevant for tweet-based AM given that tweets are assumed to contain argumentation of somewhat low quality. Tweets containing sources for factual information, however, are seen as more sophisticated. After having classified their tweet data set into argumentative and non-argumentative tweets, additional classification models were trained on the same features to separate factual from opinionated tweets. An LR model trained on the whole feature set yielded best results (F1: 0.80). Finally, in a third step the authors conducted source identification. This task was approached by using string matching and named entity recognition, as the corpus of argumentative factual tweets was too small for training an ML model. If common news agencies were found in a tweet or if, alternatively, an organisation or person was named, whose abstract on DBpedia[9] mentioned the words *news*, *newspaper* or *magazine*, a tweet was considered as containing a source. The authors obtained an F1 score of 0.67.

SS2020 also worked on evidence detection. This task was approached similarly to their work on argument and claim detection, i. e. the same feature sets and ML models were used. Applying classification on their full tweet annotations yielded an F1 score of 0.67, which is a reduction compared to the respective argument and claim detection results. However, applying sequence labeling to their ADU annotations yielded an F1 score of 0.75, which is comparable to the score achieved in ADU annotation-based argument detection. Importantly, this score surpasses the results achieved in claim detection.

To summarize, evidence detection was identified as a core task in tweet-based AM, as it is linked to the important and related task of quality evaluation. Also social media-specific evidence types were defined. As in general argument and claim detection, current approaches rely on supervised classification and sequence labeling. Specifically, SVM, LR, XGBoost, and CRF models were used.

4.4 Relation detection

While the majority of work concentrates on the identification of argument components, first work exists that examines argument relations, i. e. support or attack. This step is needed, for instance, if the AM pipeline is supposed to involve the analysis of argument structure between tweets.

BCV2016b investigated relation detection as the third step of their AM pipeline. Experiments were based on the relation annotation layers of the DART corpus that depend on previously created tweet pairs. First, the authors approached the task using textual entailment [7], given the conceptual proximity between support/entailment and attack/contradiction, respectively. Recall that the DART data was annotated both for support/attack relations and entailment. Textual entailment was predicted using the Excitement Open Platform [19], however, results were not promising (F1 (support): 0.17; F1 (attack): 0.00). The authors associated this with a mismatch between the classes in the two annotation layers and a high degree of unrelated

[9] https://www.dbpedia.org/

tweet pairs. In a next step, BCV2016b applied a neural classification approach based on a Long Short Term Memory (LSTM)-driven [15] encoder-decoder network. Still, the results were unsatisfying (F1 (support): 0.20; F1 (attack): 0.16), which was attributed to the difficulty of the task. The authors pointed out the challenge of pairing tweets sensibly and dealing with the complexity of some tweet's content.

4.5 Graph building

Modelling a full discussion on Twitter is one possible target application of tweet-based AM. To accomplish this aim, building argument graphs is a crucial step. BCV2016b explored this task by treating argumentative tweets as nodes in the graph linked by edges representing support/attack relations. Generally, such a graph consists of many subgraphs, the smallest of which are based on the formerly created tweet pairs.

Related work was presented by SS2020 who utilized context information via reply relationships between tweets. Although this study was not designed with graph building in mind, future work may focus on employing Twitter conversation information in addition to relations based on semantic proximity (as in BCV2016b) in order to derive a more complete picture of a given discussion.

Interim conclusion

In line with the annotation objectives presented in Section 3, previous work focused on the detection of the core components of argumentation. However, most studies refrained from specifically detecting claims and instead concentrated on identifying argumentative tweets, in general, before continuing with a subsequent task (e. g. evidence detection) (AB2016, BCV2016b, DCV2017). So far, only little work has been done on relation detection (BCV2016b) and ADU level argument component detection (SS2020).

5 Stance detection and AM

While AM consists of many sub-disciplines, some of which have been discussed in the last sections, it is also linked to a number of related NLP tasks (e. g. stance detection and sentiment analysis[10]). This results from the fact that these tasks all revolve around the question how to detect a standpoint towards a target. In this section we will discuss work on one related NLP task, namely stance detection. Stance detection may be defined as the task of detecting a person's position with respect to a given target, which may be a controversial topic or issue like climate change [28]. In contrast with AM it is not primarily interested in detecting constellations of reasons and counter-considerations for claims/positions. In that sense, stance detection can be seen as a surrogate or as a supplement for AM. We will discuss evidence that this benefit can further work in both directions. Results of the papers discussed in this section can be found in Table 7.

Table 7: Stance Detection Results.

Paper	Target	F1
ASB2017	Topic	0.94
	Stance	0.90
SS2019	debate (n-gram)	0.70
	debate (USE)	0.78
WZ2016	explicit	0.78–0.95
	debate (n-gram)	0.66
	debate (oracle)	0.88

Relevant work on the intersection of AM and stance detection was published by WZ2016. In particular, they focused on the issue of implicitness in tweet-based argumentation and a potential solution via a stance detection approach. The study was based on the hypothesis that an implicit claim is always linked to the stance towards the overall debate target (e. g. *Atheism*). While this *debate stance* tends to be implicit as well, it may be practical to approach it (and thereby also the implicit claim) via other stances, explicitly mentioned in the data. Example (9) shows a tweet explicitly expressing a positive stance towards *Christianity*. Assuming that this tweet is part of a debate about *Atheism*, we may conclude that the author of the tweet implicitly expresses a negative stance towards the debate target.

(9) Bible: infidels are going to hell!

In this work, WZ2016 utilized English tweet data published for the shared task on automated stance detection (Subtask a), which was associated with SemEval 2016 [21]. Specifically, the authors used the sub-corpus on the target *Atheism*. In a first step, WZ2016 created stance annotations with respect to the debate target, i. e. Atheism, and additionally to a set of derived atheism-related targets

[10] Due to space limitations we will not discuss sentiment analysis in this work.

(e. g. *supernatural power*, *Christianity*, *Islam* etc.). While the stance towards the debate target was allowed to be implicit, annotators were instructed to only annotate stance towards the derived targets if textual evidence was given. A Fleiss' ϰ score of 0.63 was achieved for the joint annotations of debate and explicit stances.

Subsequently, WZ2016 trained a linear SVM on word and character n-grams. F1 scores of 0.78–0.95 were yielded for the most common explicit targets.[11] For the debate target an F1 score of 0.66 was achieved. This score was compared to the results of an oracle model which has been trained directly on the explicit stance annotations (F1: 0.88) to see the potential of the approach. Given the substantial difference between both results, the authors concluded that a more successful detection of explicit stances is needed.

Building on WZ2016, [24] (henceforth SS2019) improved on these results by experimenting with different kinds of word and sentence embeddings. As SS2019 utilized the same annotations and a similar classification model as WZ2016, the results are comparable. Best results (F1: 0.78) were achieved using a pre-trained language-agnostic Universal Sentence Encoder (USE) model [9]. Interestingly, a model trained on GloVe embeddings which had been pretrained specifically on Twitter data yielded worse results (F1: 0.60). This study showed that the results of WZ2016's oracle condition can already be approximated by choosing a different feature set. Further improvements may be achieved by adjusting the model architecture or the modelling of the explicit stances.

While WZ2016 made use of stance detection to approximate AM, [2] (henceforth ASB2017) reversed the direction of inference and attempted to improve stance detection by employing the argument annotation layers originally created in AB2016. Thus, this work lends support to the hypothesis, that AM and stance detection can benefit from each other.

Recall that the corpus created in AB2016 included 3000 tweets on the Apple/FBI encryption debate of 2016. Two annotation layers were created: general argumentativeness and evidence type. ASB2017 created two additional stance-related layers: 1) the target, i. e. *national security*, *individual privacy*, *other*, or *irrelevant*; 2) the stance towards the target, i. e. *favor*, *against*, or *neutral*. The authors achieved Cohen's ϰ scores of 0.70 and 0.64 for target and stance annotation, respectively. While these scores are promising, one also has to consider the high degree of class imbalance in both annotation layers, which could bias the classifier. To counteract this eventuality, the authors applied under- and oversampling.

ASB2017 utilized four types of features: lexical (e. g. unigrams), syntactic (e. g. number of POS occurrences), Twitter-related (e. g. hashtags), and argumentation (i. e. the annotations created in AB2016). In addition, they chose to experiment with NB, SVM and DT models. First, models were trained to detect a tweet's target, i. e. *national security* and *individual privacy*. Best results were achieved by a DT model trained on all features (F1: 0.94). Second, the authors conducted stance classification. Different features were tried, out of which a set of lexical (unigrams and bigrams) and argumentative (argumentativeness and evidence type) features in combination with an SVM performed best (F1: 0.90). Thus, ASB2017 provided evidence that stance detection can be improved by using argumentative features.

6 Conclusion and outlook

In this paper, we presented a critical survey of the state of the art in tweet-based AM. We showed that previous work developed new approaches to argument modelling that take into account certain Twitter characteristics (topic vs conversation-based discussions). In this context several tweet corpora were annotated for argumentation. We further discussed progress on the tasks of (general) argument, claim, evidence, and relation detection. Previous studies defined these tasks as supervised classification problems or, less frequently, as sequence labeling problems, depending on the unit of annotation (full tweet vs ADU). Given that tweet corpora tend to contain a notable amount of non-argumentative content, filtering for argumentative tweets by using binary classification was a common starting point. While claim detection also has been investigated, evidence detection proved to be the more frequent task of interest. Finally, we focused on the intersection of AM and stance detection. Promising results indicate that both disciplines are interconnected and can be fruitfully employed to advance each other. We conclude that important first steps have been taken in tweet-based AM. However, several areas of further development can be identified:

1. Advance AM approaches by employing deep learning and neural network techniques.
2. Integrate topic-based and conversation-based discussion.

[11] Importantly, classes of the explicit targets are rather unbalanced, which is reflected in the different F1 scores.

3. Extend research on tweet-based AM to other languages.

So far, AM on Twitter is mainly focused on traditional ML techniques (e. g. SVM-based supervised classification). Recent progress in the development of neural network architectures and approaches may be utilized to advance this field further as well. Given that neural networks tend to require large amounts of training data, we argue that progress in this regards has been hindered due to the restricted amount of tweet corpora annotated for argumentation. As corpus annotation is an expensive task, it remains unclear if this issue will be solved in the near future. One partial solution may be annotations with so-called *weak labels*, i. e. labels that are easily created (e. g. via a heuristic) but are characterized by a lack of quality. For instance, one may annotate tweets with certain hashtags as argumentative by using a simple string matching approach. Previous research has shown that blending high quality annotations with weakly labelled data can improve training of neural networks [26].

Second, we consider it relevant to integrate both topic-based (BCV2016a) and conversation-based (SS2020) discussions into one model of tweet-based argumentation. This points to a general issue of argument annotation in tweets. While both independent and reply tweeting can be used to participate in a discourse, each type has its consequences for mining argumentation. Whenever researchers decide to focus on one type, they are bound to miss the full picture. If, however, a full tweet-based AM pipeline is to be developed, i. e. including tasks like relation detection and graph building, all types of cross-tweet relations should be considered. Indeed, this would include a third and so-far unstudied type of *conversation*: the thread. A Twitter thread refers to a series of linked tweets posted by one person. Such a constellation has the potential to be argumentative as the user makes an effort to convey their statement in a more extensive way.

The majority of previous research on tweet-based AM has concentrated on English data. However, while certainly most tweets are written in English, other languages are frequently used as well, including Japanese, Spanish, and Malay.[12] A comprehensive approach to AM on Twitter should also imply research on these understudied languages. Of course, this is not feasible for individual AM practitioners, given that sophisticated knowledge of the to-be-investigated language is needed for analysing argumentation. Furthermore, the lack of annotated corpora is especially severe for these languages. Alternatively, it may be practical to partly approach this issue by employing language-agnostic text embeddings. Previous research on this feature type could assist with bridging from one language to another [12].

[12] For statistics on used languages on Twitter, see: https://www.statista.com/statistics/267129/most-used-languages-on-twitter/

References

1. Aseel Addawood and Bashir Masooda. “What is your evidence?” A study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11, Berlin, Germany, August 2016. Association for Computational Linguistics.
2. Aseel Addawood, Jodi Schneider, and Bashir Masooda. Stance classification of twitter debates: The encryption debate as a use case. In *Proceedings of the 8th International Conference on Social Media & Society*, New York, NY, USA, 2017. Association for Computing Machinery.
3. Berfin Aktaş, Veronika Solopova, Annalena Kohnert, and Manfred Stede. Adapting coreference resolution to Twitter conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2454–2460, Online, November 2020. Association for Computational Linguistics.
4. Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, December 2008.
5. Tom Bosc, Elena Cabrio, and Serena Villata. DART: A dataset of arguments and their relations on Twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1258–1263, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).
6. Tom Bosc, Elena Cabrio, and Serena Villata. Tweeties squabbling: Positive and negative results in applying argument mining on social media. In *Computational Models of Argument – Proceedings of COMMA 2016*, Potsdam, Germany, 12–16 September, 2016, volume 287 of *Frontiers in Artificial Intelligence and Applications*, pages 21–32. IOS Press, 2016.
7. Elena Cabrio and Serena Villata. A natural language bipolar argumentation approach to support users in online debate interactions. *Argument & Computation*, 4(3):209–230, 2013.
8. Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16*, pages 785–794, New York, NY, USA, 2016. Association for Computing Machinery.
9. Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Learning cross-lingual sentence representations via a multi-task dual-encoder model. *CoRR*, 1810.12836, 2018.
10. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

11. Mihai Dusmanu, Elena Cabrio, and Serena Villata. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
12. Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding, 2020.
13. Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. Argument extraction from news, blogs, and social media. In *Artificial Intelligence: Methods and Applications*, pages 287–299, Cham, 2014. Springer International Publishing.
14. Kathrin Grosse, Carlos Iván Chesñevar, and Ana Gabriela Maguitman. An argument-based approach to mining opinions from twitter. In *AT*, volume 918 of *CEUR Workshop Proceedings*, pages 408–422. CEUR-WS.org, 2012.
15. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.
16. John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.
17. Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and N. Slonim. Context dependent claim detection. In *COLING*, 2014.
18. Clare Llewellyn, Claire Grover, Jon Oberlander, and Ewan Klein. Re-using an argument corpus to aid in the curation of social media collections. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 462–468, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).
19. Bernardo Magnini, Roberto Zanoli, Ido Dagan, Kathrin Eichler, Guenter Neumann, Tae-Gil Noh, Sebastian Pado, Asher Stern, and Omer Levy. The excitement open platform for textual inferences. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 43–48, Baltimore, Maryland, June 2014. Association for Computational Linguistics.
20. Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law, ICAIL '07*, pages 225–230, New York, NY, USA, 2007. ACM.
21. Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California, June 2016. Association for Computational Linguistics.
22. Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):1–31, January 2013.
23. Rob Procter, Farida Vis, and Alex Voss. Reading the riots on twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*, 16(3):197–214, 2013.
24. Robin Schaefer and Manfred Stede. Improving implicit stance classification in tweets using word and sentence embeddings. In *KI 2019: Advances in Artificial Intelligence*, pages 299–307, Cham, 2019. Springer International Publishing.
25. Robin Schaefer and Manfred Stede. Annotation and detection of arguments in tweets. In *Proceedings of the 7th Workshop on Argument Mining*, pages 53–58, Online, December 2020. Association for Computational Linguistics.
26. Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia, July 2018. Association for Computational Linguistics.
27. Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar, October 2014. Association for Computational Linguistics.
28. Manfred Stede and Jodi Schneider. *Argumentation Mining*, volume 40 of *Synthesis Lectures in Human Language Technology*. Morgan & Claypool, 2018.
29. Giuseppe A. Veltri and Dimitrinka Atanasova. Climate change on twitter: Content, media ecology and information sharing behaviour. *Public Understanding of Science*, 26(6):721–737, 2017. PMID: 26612869.
30. Jan Šnajder. Social media argumentation mining: The quest for deliberateness in raucousness, 2016.
31. Michael Wojatzki and Torsten Zesch. Stance-based argument mining – modeling implicit argumentation using stance. In *Proceedings of the KONVENS*, pages 313–322, 2016.
32. Sarita Yardi and Danah Boyd. Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology & Society*, 30(5):316–327, 2010.

Bionotes

Robin Schaefer
University of Potsdam, Dept. Linguistics,
D-14476 Potsdam, Germany
robin.schaefer@uni-potsdam.de

Robin Schaefer is a Ph.D. student in the Applied Computational Linguistics group (headed by Manfred Stede) at the University of Potsdam, Germany. His thesis concentrates on the detection of argumentation structure in text with a special focus on Twitter. He is also affiliated with the Qurator project, based at the State Library Berlin, Germany. Here, he focuses on OCR post-correction. He obtained his M.Sc. in Psycholinguistics from the University of Potsdam in 2018.

Manfred Stede
University of Potsdam, Dept. Linguistics,
D-14476 Potsdam, Germany
stede@uni-potsdam.de

Manfred Stede is a professor of Applied Computational Linguistics at the University of Potsdam, Germany. He obtained his Ph.D. in Computer Science from the University of Toronto in 1996 with a thesis on language generation. After working for five years in a machine translation project at TU Berlin, he moved to Potsdam in 2001, where his interests shifted to text analysis. He conducted research projects on applications like information extraction and discourse parsing, and on more theoretical matters like the semantics and pragmatics of discourse relations and connectives. Several years ago, he proceeded to work on approaches to deriving argumentation structure trees from short texts, and on various other aspects of argumentation mining. He (co-)authored four monographs, including *Argumentation Mining* (with Jodi Schneider), and around 150 peer-reviewed papers in journals, conferences and workshops.